SPEECH SYNTHESIS BY RULE:

AN ACOUSTIC DOMAIN APPROACH

by

LAWRENCE RICHARD RABINER

S.B., S.M., Massachusetts Institute of Technology
1964

SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June, 1967

Signature of Author _____
Department of Electrical Engineering, May 12, 1967

Certified by _____
Thesis Supervisor

Accepted by _____
Chairman, Departmental Committee on Graduate Students

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUN 1967		**00-06-1967 to 00-06-1967**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Speech Synthesis by Rule: An Acoustic Domain Approach	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Massachusetts Institute of Technology,77 Massachusetts Avenue,Cambridge,MA,02139**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**214**	

SPEECH SYNTHESIS BY RULE:
AN ACOUSTIC DOMAIN APPROACH

by

LAWRENCE RICHARD RABINER

Submitted to the Department of Electrical Engineering on May 12, 1967 in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

ABSTRACT

A novel approach to speech synthesis by rule is presented and evaluated. A discrete set of input symbols is converted to continuous control signals for driving a serial terminal analog speech synthesizer. The synthesizer converts the control signals to a continuous acoustic output. The synthesizer was simulated on a digital computer.

The input set is a linguistic description of the utterance. It includes information about which phonemes are present, which vowels are stressed, and where sentence and word boundaries and pauses occur. The input also specifies whether or not the utterance is a yes-no question.

The process of converting from the discrete input to the control signals is called the synthesis strategy. In this strategy, each phoneme has a characterization which contains relevant information about the acoustic parameters of the phoneme. Control signals are generated by proceeding between acoustic parameter values appropriate for the phonemes in the input string. Formant transitions between phonemes are generated by solutions to a mathematical equation. Formant transition time constants are determined from matrices of time constants and are functions only of the two phonemes between which the transition is being made. Formants move semi-independently of each other. For each phoneme a set of frequency regions around the formant target values is specified. These frequency regions, in general, determine the timing of events. A transition from phoneme A to phoneme B cannot begin until all formants are within the frequency regions of phoneme A. If phoneme A is a stressed vowel, a specified duration of steady state is generated before the transition to phoneme B begins. Otherwise the transition to phoneme B begins as soon as the formants are within the frequency regions of phoneme A.

The remaining synthesizer control parameters (except fundamental frequency) are generated from the formant frequency contours. These parameters initiate and terminate motion at characteristic times during formant transitions.

The fundamental frequency contour is generated from information about where sentence boundaries occur, which vowels are stressed, and whether a consonant is voiced or voiceless. A model for predicting fundamental frequency from this information is postulated. The controlling variables are subglottal pressure and laryngeal tension. Archetypal contours for these variables are postulated and these contours are modified by the content of the utterance.

Consonant intelligibility tests have indicated a high degree of intelligibility for synthetic consonants in both prestressed and post-stressed word position. Sentence intelligibility tests have yielded high intelligibility scores for both simple and complex sentences.

Many problems are as yet unsolved. The percent intelligibility scores for certain consonants are low. Many of the synthetic sentences were judged unnatural sounding. Rules for modifying the fundamental frequency contour and/or the timing of events, based on the syntactic structure of the sentence, are needed.

Thesis Supervisor: Kenneth N. Stevens
Title: Professor of Electrical Engineering

DEDICATION

This thesis is dedicated to my parents who have always encouraged the pursuit of knowledge.

ACKNOWLEDGEMENT

The author wishes to express his indebtedness to his supervisor, Professor Kenneth N. Stevens, for his guidance and assistance throughout this thesis work. His excellent advice and continued interest in this thesis were a constant source of inspiration. To his thesis readers Professors Dennis Klatt, and William L. Henke, the author wishes to express his appreciation for the many valuable suggestions they made.

The computer facilities of Bell Telephone Laboratories were used throughout this thesis work. The author wishes to thank Dr. James L. Flanagan for permission to use these facilities. Further thanks are due Misses Linda L. Gibson and Virginia Sachse for valuable assistance in running computer programs, processing data, and in the preparation of this manuscript.

This work was done in the Speech Communications group of the Research Laboratory of Electronics. The work of this group is supported by a grant from the National Institutes of Health, and a contract from the U.S. Air Force Office of Aerospace Research.

Part of this work was done while the author was on a summer assignment at Bell Telephone Laboratories, Murray Hill, in the Department of Speech and Auditory Research.

The author wishes to thank the National Science Foundation for providing financial support in the form of a Cooperative Fellowship from September 1964 to June 1966 and a Graduate Fellowship from September 1966 to June 1967.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

INTRODUCTION

1.1 Introduction to Synthesis by Rule

In recent years much interest in studying speech synthesis has been aroused. The reasons for this are twofold. The primary reason has been the availability of high speed digital computers and the ease of man-machine communications by means of graphical input-output devices, teletypes, and display consoles. These computers have enabled researchers to study and evaluate ideas that would be unfeasible to study in any other way. The second reason for interest in speech synthesis has been the increased understanding of the process of speech production. X-ray data, air flow measurements, and acoustic analyses of speech, among other techniques, have led to greater understanding of the functions and workings of the vocal tract mechanisms, and of such influences as source-system interactions in the production of speech.

Research in speech synthesis has proceeded along two lines for the most part. Along the first line researchers have tried to discover the significant cues, or parameters, which enable listeners to perceive a certain sound or sensation. Some machine (possibly a computer) had to be used to synthesize the speech, and this machine periodically required information as to parameter values for its control signals. These parameters had to be specified by the researcher. Along the second line researchers have tried to synthesize speech from a linguistic description of the utterance. Here the job of the researcher has been to suitably code this linguistic description into the necessary control information for his speech synthesizer. This process is called "Speech Synthesis

by Rule." It is this aspect of speech synthesis that we will be concerned with in this thesis.

The act of speaking is undoubtedly a complicated process. Only man, among all the living creatures, has been given the ability and the tools to converse in a meaningful way. There are many levels of organization between the highest level, that of having some thought one wishes to express, and the final acoustic output. The highest levels involve syntactic, semantic, and phonological components of the language. In this thesis, we will not be primarily concerned with what occurs in these higher levels. We will assume that the original thought has been coded into a sequence of discrete units. It is generally agreed that these units are the phonemes (Halle, 1964). The information from the phonemes is coded into a set of control signals for the articulatory structures (teeth, lips, tongue, etc.), and then the articulators move, thus generating the continuous acoustic speech signal.

In this thesis we are concerned with the conversion between the discrete data at the level of phonemes to a continuous acoustic output. What occurs in the articulatory domain will be of interest as a guiding principle, and to gain insight into how and what events should occur - but we will not be concerned with such problems as a physical description of tongue position, or velocity along points of the tongue, etc. (Speech synthesis by modelling the speech generator structures of the vocal tract has been investigated by Henke, 1966.)

The choice of the acoustic domain for our rules for speech synthesis deserves some comments. An acoustic description of speech is inherently simpler than an articulatory description because of the smaller number of parameters. There is a wealth of information available as to

values of these parameters in a variety of situations. Rules can be written and an analysis of the speech can be made, thus providing valuable feedback as to the successes and failures of the rules. Rules can be written to account for contextual and coarticulational features of speech. Finally, speech synthesized from acoustic parameters, where the parameters have been extracted by eye, has proved to be of high quality - thus justifying and giving motivation to attempts at writing rules to give equally high quality speech.

The basic shortcoming of writing rules for synthesizing speech from acoustic parameters is that some of the rules are ad hoc, having little physiological justification other than they work well this way. Hopefully the majority of the rules will have a sound basis for their existence and hence will be of lasting value.

1.2 Outline of Synthesis Strategy

Our general approach has been to formulate a framework for a synthesis strategy from the body of available data in the literature. Whenever necessary we have generated our own data to test or expand on existing data. Once the framework was established tests were run to evaluate parameters of our model. Changes were made and evaluated by subsequent tests.

The strategy we have used in this thesis is basically as follows. Each phoneme has a characterization which contains relevant information about target values of the acoustic parameters of the phoneme. (The acoustic parameters are those used to drive a speech synthesizer.) The input to the strategy contains information about which phonemes are present, where stressed vowels occur, where word boundaries occur, and where pauses are to be inserted. Control signals are generated by proceeding between

acoustic parameter target values appropriate to the phonemes of the input string. Formant transitions between phonemes are generated by the solution to a mathematical equation. Transition time constants are determined from a matrix of time constants. Formants move semi-independently of each other. For each phoneme a set of frequency regions around the formant target values is specified. These frequency regions, in general, determine the timing of events. A transition from phoneme A to phoneme B cannot begin until all formants are within the frequency regions of phoneme A. If phoneme A is a stressed vowel, a specified duration of steady state is generated before the transition to phoneme B begins. Otherwise the transition to phoneme B begins as soon as the formants are within the frequency regions of phoneme A.

The fundamental frequency contour is generated from information about which vowels are stressed, whether consonants are voiced or voiceless, and where the sentence boundaries lay. A model for predicting fundamental frequency from this information is postulated. The controlling variables are subglottal pressure and laryngeal tension. Archetypal contours for these variables are postulated and these contours are modified by the content of the utterance.

1.3 Outline of the Report

Chapter two provides a discussion of the results of previous synthesis by rule schemes. The emphasis is placed on determining both the major difficulties of these schemes and the major successes. A detailed description of the terminal analog synthesizer used in this thesis is presented in Chapter three. A short discussion of the general characteristics of a terminal analog synthesizer is also included in Chapter three. Chapter four is intended to provide a bridge between the

results of experiments on speech analysis, synthesis, and perception, and the specific synthesis strategy we have adopted. Quasi-static characterizations of the phonemes are presented. These characterizations correspond to optimal synthesizer configurations for steady state representations of the phonemes. Chapter five presents the details of the synthesis strategy. The general attributes of an acoustic domain approach to synthesis by rule are presented along with the details of our specific implementation. A quantitative evaluation of our strategy is presented in Chapter six. Consonant intelligibility and word intelligibility test results are given and evaluated. Chapter seven presents a summary of the results, an overall evaluation of our technique, and suggestions for future research.

CHAPTER TWO

PREVIOUS ATTEMPTS AT SYNTHESIS BY RULE

2.1 Introduction

In the past ten years there have been many attempts at synthesis by rule. Basically there have been two approaches to this problem. One approach involved using either a dynamic model of the vocal tract or a terminal analog synthesizer, and supplying it with a sequence of control signals to generate the speech. The second approach utilized discrete segments of human speech - connecting them together to produce the speech. These segments were supposed to contain the important information about the dynamic (transitional) and transitory properties of speech. In the following sections we will review and discuss the strong and weak points of four of these attempts. This will hopefully provide a focal point for the issues involved in synthesis by rule.

2.2 Synthesis From Segments

One of the earliest attempts at synthesis by rule was made by Peterson, Wang and Sivertsen, (1958). These investigators felt that a quantitative understanding of the dynamics of the vocal tract was limited so they studied synthesis from a library of stored waveform segments. Their prime motivation was the assumption that for synthetic speech to sound natural, the normal dynamics of speech production had to be maintained. They felt that their basic segment of speech should consist of the adjoining portions of two phonetic units. The transitional changes between phonetic units were wholly contained within a segment. Speech was synthesized by abutting these segments in a serial fashion. Thus the word seed

(#-s-i-d-#) would consist of four segments -- (#-s), (s-i), (i-d), (d-#) -- where # means silence.

There was inherently a problem with such a simple scheme and this was the need to consider such effects as intonation, duration, and stress. These workers realized that several versions of each segment were necessary to meet the diverse situations in which each segment could occur. Having made some assumptions as to the correlation between stress and duration, these workers decided that a segment would need as many as nine versions to meet all situations. This set of nine segments was called a dyad. Thus the need arose for storing about 8000 segments in order to synthesize an idiolect of American speech. This is a rather large library.

Even with such a large library, problems existed in abutting segments. Generally there was a discontinuity in the waveform at the junction and this discontinuity produced a noticeable transient. (This transient could have been eliminated.) Further there were problems with matching fundamental frequency, harmonic amplitudes and phases, formants, and overall amplitude envelopes between adjacent segments. Finally there were also unsolved problems concerning stress and duration.

Perhaps the most noteworthy aspect of this work was that it tended to discourage researchers away from speech synthesis from stored waveforms, except for certain specific applications. It motivated studies leading to quantitative descriptions of the behavior of the vocal tract during speech production, and the acoustical implications of this behavior.

2.3 Haskins Laboratories Approach to Synthesis by Rule

Workers at the Haskins Laboratories wrote rules for synthesizing speech in terms of phonemes rather than syllables. (Liberman, et al., 1959.)

The machine on which the speech was synthesized was a Pattern Playback machine. On this machine an operator hand paints a spectrogram and this schematized spectrogram is converted to sound by means of an electro-optical system. What the Pattern Playback machine essentially does is to sum the first 50 terms of a Fourier series approximation to the speech wave, keeping a constant fundamental frequency of 120 Hz.

Extensive perceptual experiments were run with speech from this synthesizer and from their experiments Liberman, et al., wrote rules for synthesizing speech from phonemes. These rules were intended to reflect generalizations about the acoustic cues for perception of the particular phoneme. They were intended to fit in a framework paralleling the articulatory frame, in that they were written in terms of the subphonemic dimensions of place, manner, and voicing. A rule involved all statements necessary to specify the acoustic cues, pertaining to one dimension, for a given class of phonemes. Each phoneme was represented by a set of subphonemic rules which specified how the dimensions participated in determining the spectrum pattern.

A basic objective of the Haskins people was that the rules be minimal, or as few in number as possible; and they be simple in structure. They further hoped that their rules would reflect perceptual cues that were independent and additive.

Without any exceptions for stress or context, Liberman, et al., needed thirty-one rules and these were distributed as follows. For consonants, nine rules were needed for place, five for manner and three for voicing. For vowels, two rules were needed for manner, and twelve rules for place. However it was clear that additional rules were necessary for

positional effects, as well as stress. Thus twelve position modifier rules were added to introduce changes in the articulatory characteristics of various phonemes in selected environments. A stress modifier rule had control over durations of vowels depending on the degree of stress.

An example of the rules for the phoneme B is as follows.*

Manner: Stop - (1) No sound at formant frequencies
(2) Burst of specified frequency and bandwidth follows silence
(3) F_1 locus is low
(4) F_2 and F_3 have virtual loci

Place: Labial - (1) F_2 and F_3 loci are specified
(2) Frequencies of buzz and hiss are specified

Voicing: Voiced - (1) Voice bar
(2) Duration of silence is specified
(3) F_1 onset is not delayed

Thus we clearly see that although the rules were subphonemically organized in an articulatory frame, they had to be converted to suitable values for painting the spectrogram. These rules show that there is a good deal of "table lookup" embodied in the procedure.

This scheme has not been specified in sufficient detail to be automated. The necessity of painting a spectrogram for each utterance is

---

* A list of the phonemes and the equivalent versions used throughout this thesis is shown in Table 2.1.

TABLE 2.1

DEFINITION OF SYMBOLS

PHONEME	IPA SYMBOL
I Y	i
I	I
E	ɛ
A E	æ
A	ɑ
UH	ʌ
U	ʊ
O O	μ
O W	ɔ
E R	ɜ˞
W	w
L	l
R	r
Y	y
B	b
D	d
G	g
P	p
T	t
K	k
M	m
N	n
N G	ŋ
F	f
T H	θ
S	s
S H	ʃ
V	v
THE	ð
Z	z
Z H	ʒ
H	h
C H	ʧ
J	j

highly undesirable. Even if this scheme of synthesis were highly successful, it could not be put to much practical application. Furthermore, using the Pattern Playback machine is an inefficient way of coding acoustic information and converting it to speech. Information as to the intensity of 50 harmonics (the spectrum envelope) had to be supplied continuously.

There were many difficulties with this method of synthesis. As was mentioned previously, detailed behavior of specific cases was sacrificed for generality, in an attempt to minimize the number of rules. For example, all formants initiated and terminated motion at the same time with no provisions for different transition times, or time delays between motion of individual formants. Further, fundamental frequency was not generated by rule. It was either kept at a constant value, or varied from external information unrelated to any rules.

The most attractive feature of this scheme is the concept of writing rules for synthesizing speech in terms of the subphonemic elements or features of the input sequence. Although the particular implementation was not very successful, the concepts behind it are of lasting value.

2.4 Kelly, Gerstman Approach to Synthesis by Rule

One of the earliest attempts at speech synthesis by rule in the acoustic domain was made by Kelly and Gerstman, (1961). The machine on which the speech was synthesized was a computer simulated resonance analog synthesizer. (For a thorough discussion of speech synthesizers see Flanagan, 1966, p. 166-209.) A block diagram of this synthesizer is shown in Fig. 2.1. Kelly and Gerstman wrote a simple set of rules to convert between phonemes and control signals for the synthesizer.

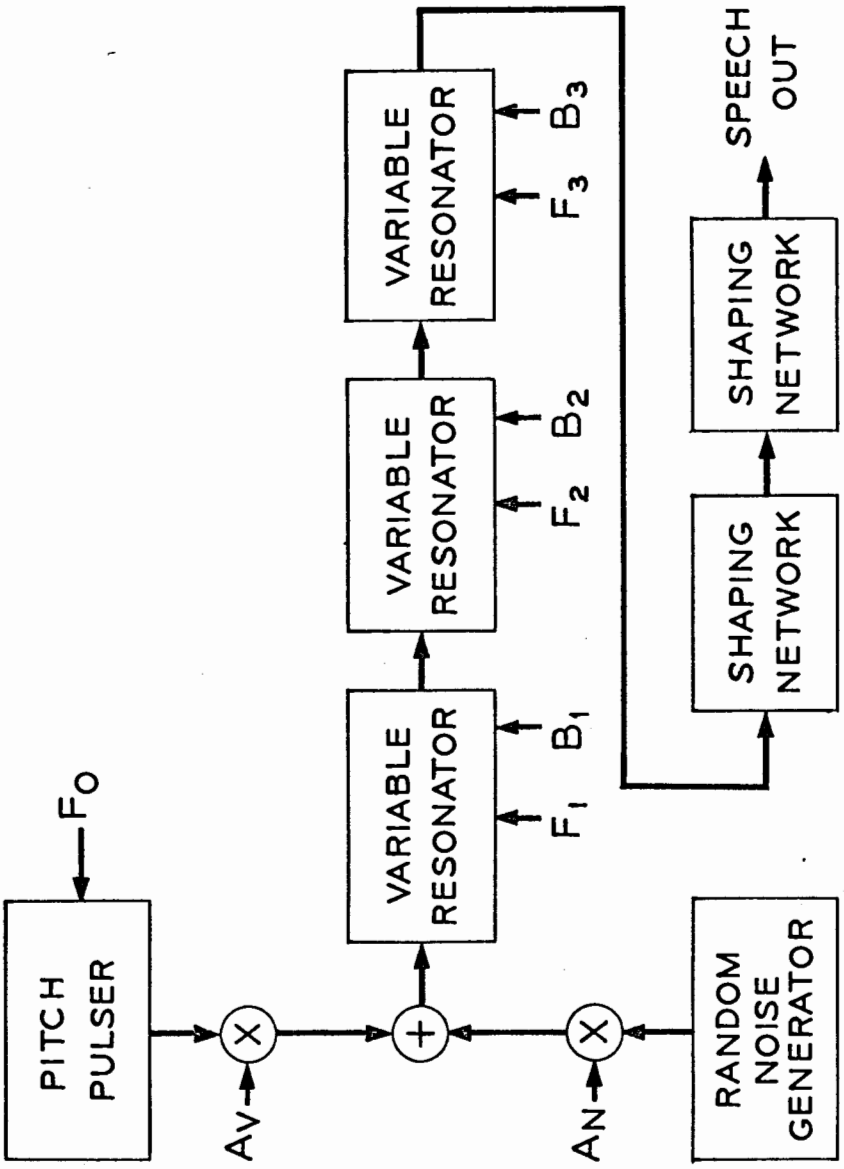


FIG. 2.1 KELLY-GERSTMAN SYNTHESIZER

As seen in Fig. 2.1 there are nine control signals. (The controls are indicated by arrows in Fig. 2.1.) These are fundamental frequency for voiced sounds, amplitudes of noise and voicing, and center frequencies and bandwidths of three variable resonators. These controls signals had to be generated periodically to produce speech. Thus the rules were a means of encoding from discrete units (phonemes) to continuous controls. There was stored for each phoneme a sequence of thirteen quantities. Nine of these quantities specified steady state or target values of the nine synthesizer control signals. The remaining four dealt with timing and transition shaping. These four parameters were steady state duration (T_{ss}), transition time into steady state (T_{IN}), transition time out of steady state (T_{OUT}), and consonant-vowel character of the phoneme. The control signals were generated by proceeding from target values of control signals for one phoneme to those for the following phoneme. Thus each control signal was a time sequence of steady values followed by transitions. The transitions between adjacent vowels or consonants were linear, whereas parabolic transitions were used between consonant-vowel and vowel-consonant pairs. The time duration of the transition was the sum of T_{OUT} for the first phoneme and T_{IN} for the second phoneme. Each transition was followed by a duration of steady state appropriate for the phoneme into which the transition went.*

Fundamental frequency was not generated by rule, but instead left as a parameter to be generated by any method desired by the operator.

---

* All relevant information in this section has been found in U.S. Government patent #3,158,685.

Further, there was provision in the rules for the insertion of a burst following a voiceless stop consonant.

This method of speech synthesis, although simple in nature, and hence attractive, is too restrictive to be able to model the dynamics of speech. Kelly and Gerstman, realizing the dynamic character of speech, tried to capture it entirely in the transitional periods of the control signals. These transitions were always followed by static periods (of non-zero duration) when no control signals were changing. Speech cannot be modeled in this way. Examination of spectrograms shows that often no steady state exists for many phonemes in an utterance.

To be more specific, the restrictive features of the Kelly-Gerstman scheme are as follows.

(1) Formants move with equal transition times. In most transitions (especially from consonant to vowel, and vowel to consonant), formant one (F1) moves more rapidly than formant two (F2) or formant three (F3). A clear example of this is seen in Fig. 2.2. Here we have the cluster IY-N-A. The transition of F1 from N to A occurs very rapidly (in many cases, a discontinuity in formant value occurs), while both F2 and F3 move much slower.

(2) Formants initiate motion to new targets at a common point in time. In many transitions into consonants formant two, and often formant three, initiates motion to new targets long before formant one moves. A clear example is seen in Fig. 2.3 for the utterance AE-B-A. Here, in the transition between AE and B we clearly see F2 moving before F1. This delay of formant motion has often been observed on spectrograms of real speech.

Figure 2.2 Wideband spectrogram of natural utterance IY-N-UH.

VOICEPRINT LABORATORIES - P. O. BOX 835 - SOMERVILLE, NEW JERSEY

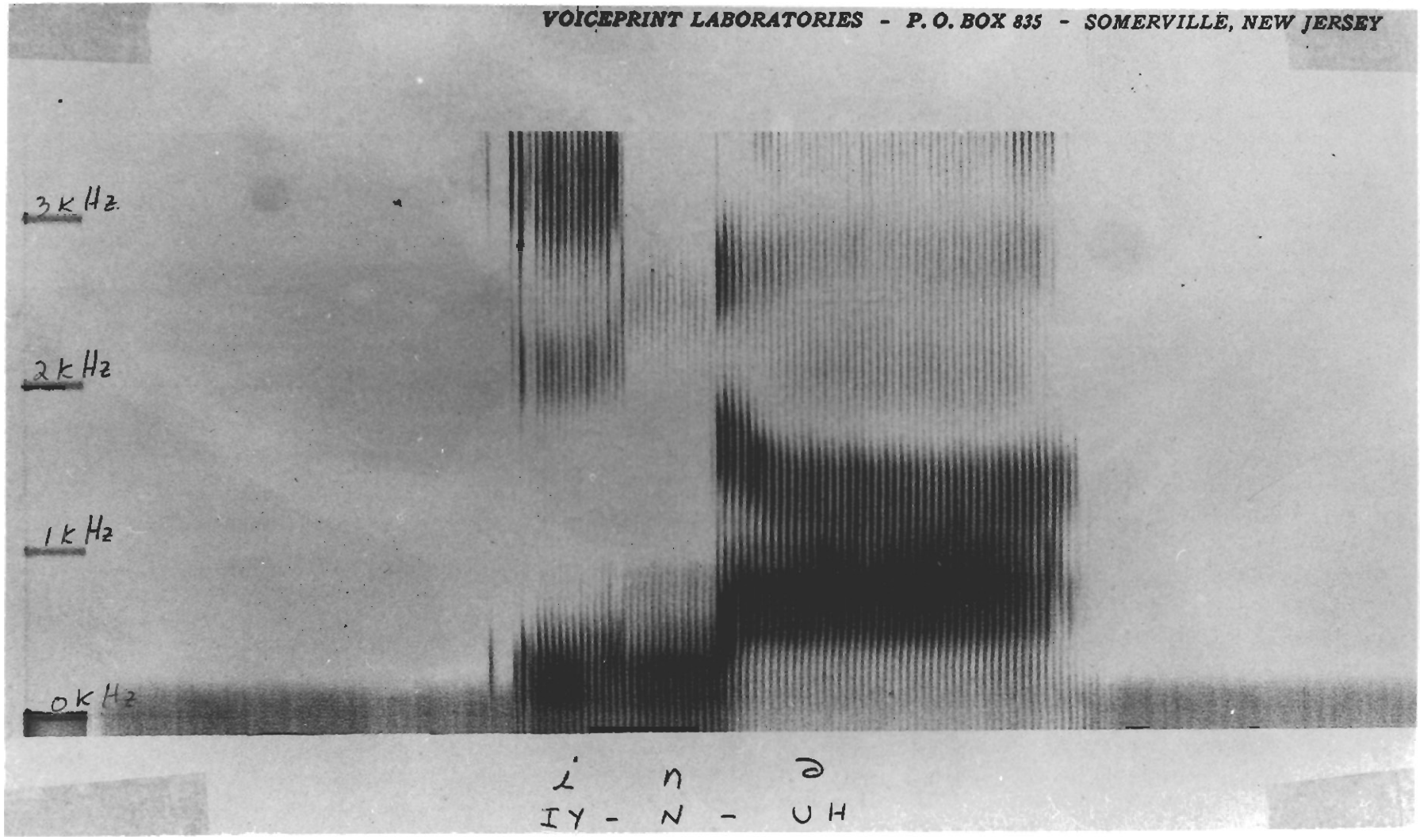
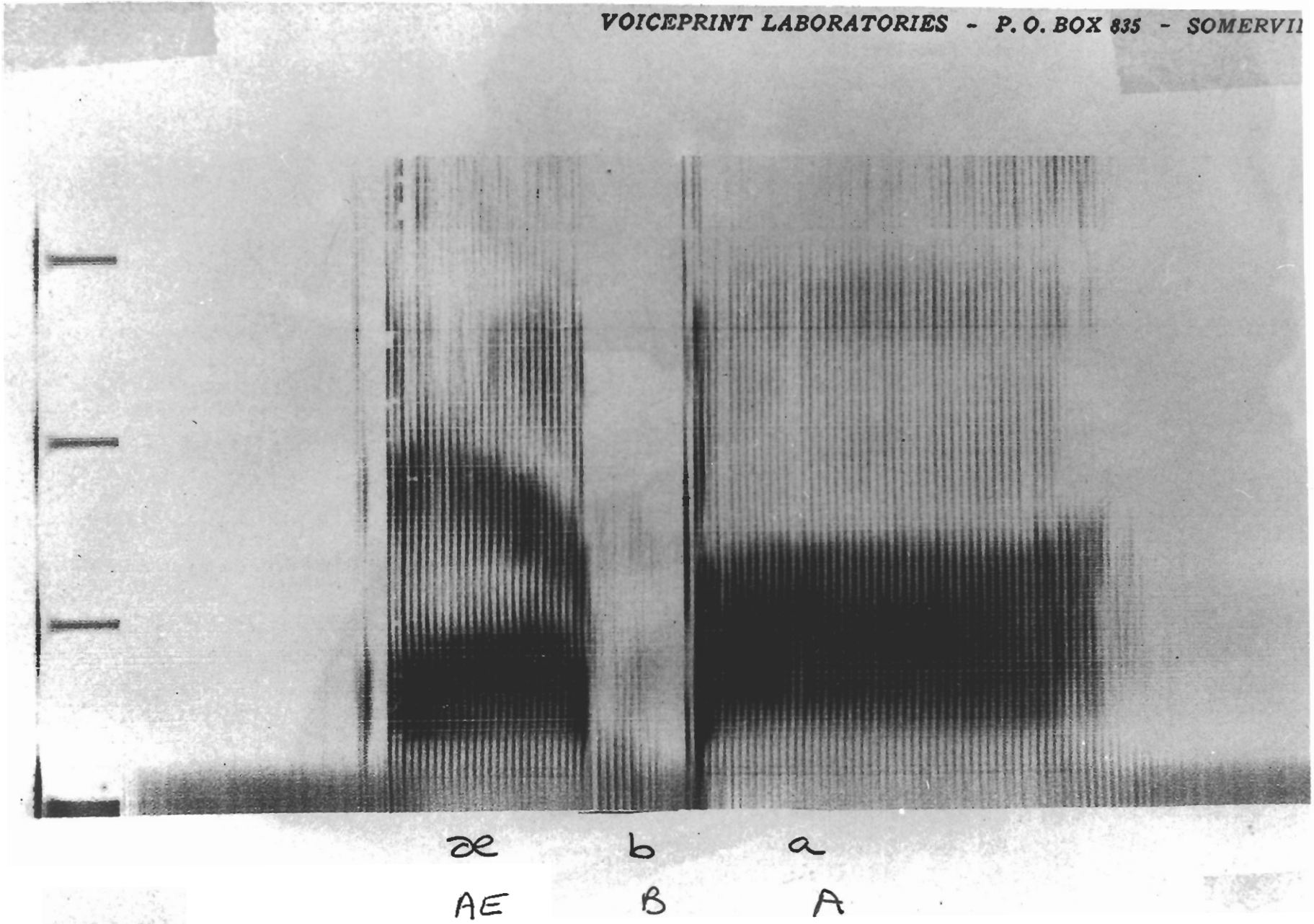


Figure 2.3 Wideband spectrogram of natural utterance AE-B-A.


VOICEPRINT LABORATORIES - P. O. BOX 835 - SOMERVI
æ b a
AE B A

(3) All formants have equal steady states, and the steady state duration of the phoneme is independent of context. As was mentioned previously, speech is not a succession of steady states connected by appropriate transitions. Unstressed vowels often exhibit no steady state. Further it is known that the duration of stressed vowels is an important cue in distinguishing between phoneme pairs such as voiced and voiceless stops which follow the vowel. Also the duration of a vowel is a strong cue as to whether it is stressed or not. The Kelly-Gerstman scheme could either give stressed or unstressed vowels - not both.

(4) The transition duration is influenced equally by both phonemes as an unweighted sum of two individual transitions. To see that this is restrictive consider the following hypothetical case. Transition time (as measured from spectrograms of real speech) from phoneme one (PH1) into phoneme two (PH2) is short; whereas from PH1 to phoneme three (PH3) it is long. Using this scheme, one would be forced to say that the transition time out of PH1 ($T_{OUT\ 1}$) is short, whereas the transition time into PH3 ($T_{IN\ 3}$) is long. Suppose we find a phoneme, say phoneme four (PH4), such that the transition time from PH4 to PH3 is short. The transition time from PH4 to PH3 is the sum of the transition times into PH3 ($T_{IN\ 3}$) and out of PH4 ($T_{OUT\ 4}$) so it must be long, and hence we cannot satisfy the requirement that it be short. Stated more concisely, the transition time T_{ij} between steady states of two phonemes i and j is:

$$T_{ij} = T_i + T_j = F(i) + F(j)$$

for the Kelly-Gerstman scheme, whereas a more general case might be:

$$T_{ij} = T_{ij} = F(i,j).$$

A specific example will serve to illustrate this point. The following results have been measured on spectrograms of the author's speech.

(a) The transition time between AE and S (as in cast - K-AE-S-T) is about 25 ms. Thus, the transition time out of AE (T_{OUT-AE}) must be less than 25 ms.

(b) The transition time between AE and M (as in lamp - L-AE-M-P) is about 150 ms. Since T_{OUT-AE} is less than 25 ms, then the transition time into M (T_{IN-M}) is more than 125 ms.

(c) The transition time between R and M (as in arms - A-R-M-Z) is about 50 ms. Since T_{IN-M} is more than 125 ms, this result cannot be accounted for using the Kelly-Gerstman scheme.

(5) Formant transitions are of a linear-parabolic nature. There is little justification, in terms of physical motion of the articulators for such a choice. Linear transitions imply sudden changes of slope of formant motion and these are unnatural and inconsistent with our concept of smooth, continuous motion of the articulators. Parabolic changes can preserve both formant value and slope at one junction, and the discontinuity at the other junction can be neglected since it can be made to occur at a point when these formants are not perceptually present (i.e., source amplitudes are set to zero).

These criticisms point up ways in which the scheme was inadequate with respect to a more general technique for synthesizing speech. An important question remains. That is just how important perceptually are the above mentioned effects? Clearly it is ridiculous to include different transition times for different formants if a human listener cannot possibly

perceive the difference. An attempt to answer this question is in the main body of this thesis.

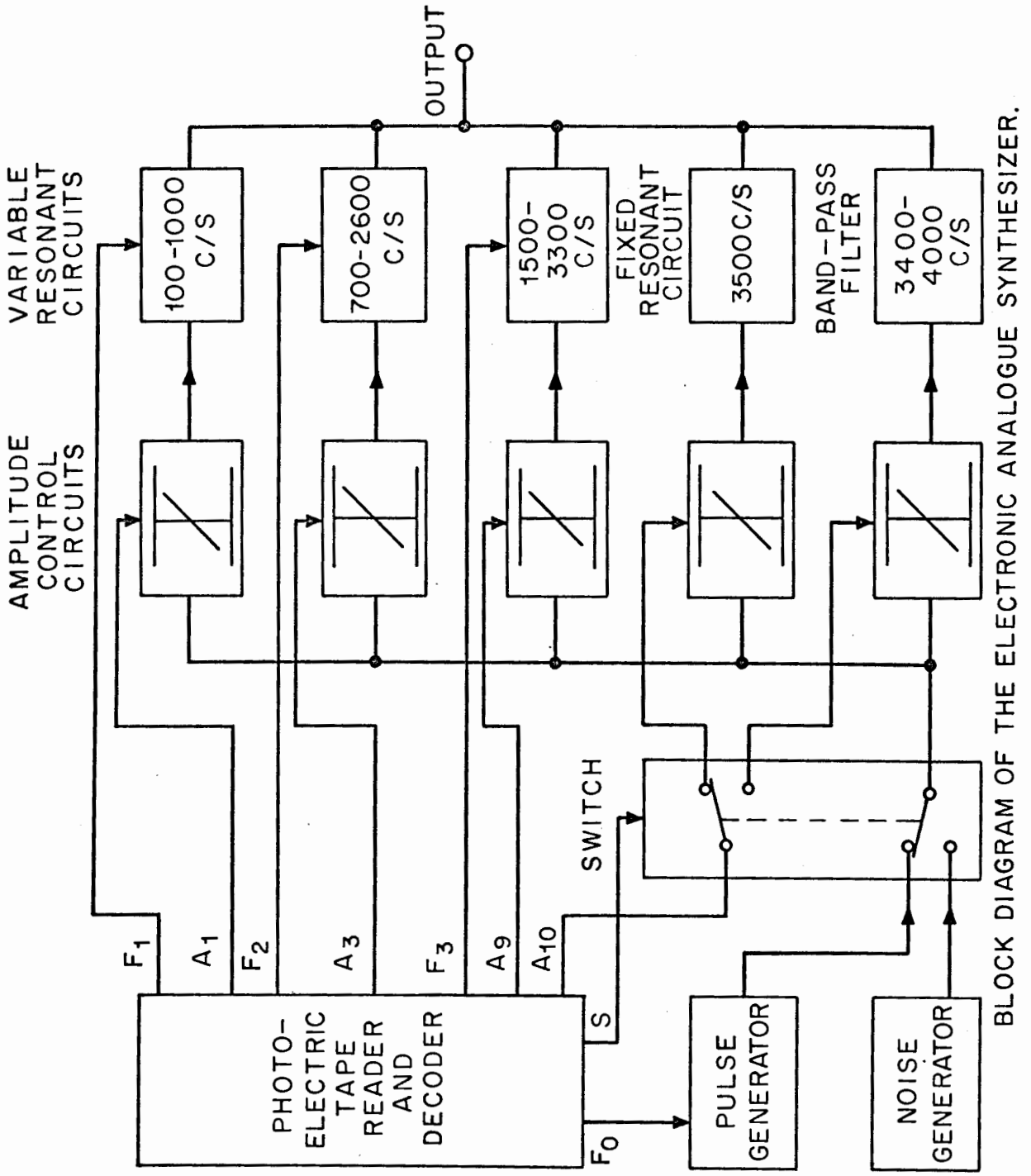
2.5 Holmes, Shearme, Mattingly Approach to Synthesis by Rule

The most successful of synthesis by rule schemes to date is that of Holmes, Shearme, and Mattingly, (1965). This was an acoustic domain approach which was in many ways similar to that of Kelly and Gerstman.

The synthesizer used here was a parallel analog synthesizer (see Fig. 2.4). As seen in Fig. 2.4 there are nine control signals including fundamental frequency. (We will discuss automatic generation of a fundamental frequency contour in Section 2.6.)

The input to the synthesis by rule scheme was a string of phonemes plus modifiers. For each phoneme there was stored in a table a series of 27 quantities which governed how to generate the eight control signals. Control signals were generated by specifying amplitudes, durations of steady state, durations of transitions, formant frequencies, weighting factors, and ranks for each phoneme. All control signals were essentially steady state values followed by transitions to new steady state values. The transitions between steady states were piecewise linear ones.

One novel, and highly successful, feature of this scheme was the influence of the phonemes on the transition times between phonemes. It is generally recognized that the character of the transition (i.e., its duration, shape) is of greater perceptual importance for some phonemes than others. For instance the vowel A can be recognized entirely from a duration of steady state whereas the consonant B exhibits no steady state whatsoever and can only be recognized from the transitions. Thus the transitions from A to B in the syllable A-B should have the character



BLOCK DIAGRAM OF THE ELECTRONIC ANALOGUE SYNTHESIZER.

FIG. 2.4 MATTINGLY, SHEARME AND HOLMES SYNTHESIZER

of B transitions. In the Holmes, et al., scheme, each phoneme was given a preference value, or weight. The character of the transition (in this case this meant the initial transition slope and the transition duration) between two phonemes was determined by the phoneme of higher weight. The weights of the phonemes are as follows. (Highest weight phonemes are highest in the list.)

a) Voiced stop consonants

b) Voiceless stop consonants

c) Voiced fricatives

d) Voiceless fricatives

e) Nasal consonants

f) W, L, R, and Y

g) Vowels and diphthongs

From this list we can see that the consonants whose perception was most determined by transition character exerted most influence on the transitions, whereas vowels, whose character was most determined by steady state durations, had no influence on transitions.

Each phoneme was characterized by a duration. When the sum of the transition times associated with the phoneme exceeded this duration, then no steady state existed. Instead, a transition towards a target was generated, but before this target was reached, a transition away from the target began. This is the commonly observed "undershoot" phenomenon.

One of the drawbacks of this scheme was the use of program modifiers. These modifiers were used to alter table entries for a phoneme in one particular position in one particular utterance. (These modifiers, derived from auditory and visual feedback, were included with the input

string for the utterance. They were not generated by rule.) In other words, these modifiers accounted for effects that the general scheme could not handle. The most generally used modifiers were for lengthening or shortening vowels and consonants in particular environments. Modifiers could also change the stratagem for synthesis of particular phonemes - such as causing a stop consonant to be unreleased. The use of modifiers not generated by rule is unacceptable for a true synthesis by rule scheme.

A further disadvantage of this scheme is the synthesizer used to produce the speech. According to Holmes et al. (1964):

"The synthesizer does, however, have certain minor limitations. The representation of the glottal volume velocity waveform provided by the pulse generator is extremely crude. Synthesis of formants in parallel imitates the behavior of the vocal tract less faithfully than would serial synthesis, and may make for poorer vowel quality (on the other hand, the independent amplitude controls permitted by the parallel arrangement are very convenient for the synthesis of consonants). ----- Finally, the forced choice between turbulent and voiced energy makes it difficult to synthesize sounds in which both forms of excitation occur, such as the voiced fricatives."

Other criticisms of this method are similar to those of the Kelly-Gerstman method. Timing is generated by specifying phoneme durations; formants move in unison at all times; all formants have identical transition times; and all transitions are linear. Without the use of modifiers, both stressed and unstressed vowels are of the same duration.

Initially fundamental frequency for this scheme was inserted by hand. At a later date a program was written by Mattingly (1966) for

automatic generation of fundamental frequency. This model will be discussed in the next section.

The speech generated by this method of synthesis by rule (albeit with the use of modifiers and a hand-inserted fundamental frequency contour) is of high quality and is clearly intelligible.

2.6 Mattingly Approach to Prosodic Features Synthesis by Rule

Synthesis of a natural sounding fundamental frequency contour has long been a problem of schemes for synthesis by rule. The most characteristic solution has been to avoid the problem entirely - that is to hand generate a fundamental frequency contour. Both intelligibility and naturalness of synthetic speech are highly dependent on the fundamental frequency contour. Hence when fundamental frequency is hand generated, the synthetic speech is greatly enhanced in quality. Furthermore when an unnatural fundamental frequency contour is used (i.e., one which could not possibly match the segmental and suprasegmental context of the utterance) the quality of the speech is greatly degraded. Rule generated fundamental frequency contours have tended to be quite bad and hence have been avoided.

A recent attempt at synthesis of a fundamental frequency contour by rules was made by Mattingly (1966). His model included three types of features. These were pausal features, intonation features, and prominence. A pausal feature was a sense-group boundary and was either a final pause, indicating the end of a sense group, or a nonfinal pause, indicating the transition between sense-groups.

The intonation features were the tunes of the sense group primarily providing information as to the syntactic structure of the utterance. In a sophistocated model the intonation features would contain information

as to the personality of the speaker, his emotional state and attitude. The three intonational features used by Mattingly were the falling tone, which indicated the end of a final clause, the fall-rise tone, which indicated the end of a nonfinal clause, and the rising tone, indicating a yes-no type of question.

Prominence was the marking of words for special emphasis. Prominence essentially implied choosing stressed syllables and modifying the fundamental frequency contour to greatly exaggerate these syllables.

Mattingly wrote rules for generating a fundamental frequency contour from these prosodic features. His contour began at a high value and fell at rates characteristic of the syllables. It continued falling until either a prominent word, or a voiceless consonant was to be generated. For a voiceless consonant the fundamental frequency stayed constant until voicing was about to begin at which point it jumped one step. (The fundamental frequency range was divided into a fixed number of equal logarithmic steps.) Following the jump fundamental frequency began its fall again.

Prominent syllables produced a three step increase in fundamental frequency. When a voiceless consonant preceded the prominent syllable, the increase was four steps and it occurred just prior to the initiation of voicing. Otherwise the three step increase was spread over the stressed syllable. Following the increase fundamental frequency again fell.

Rules were written for adjusting fundamental frequency with respect to the intonational features. These rules were ad hoc and no insight would be gained by discussing them in detail.

Mattingly's rules were a first order approximation for RP British English. They were ad hoc to a large degree. Mattingly took

no explicit account of the mechanisms by which fundamental frequency is physiologically determined. His results leave significant room for improvement.

CHAPTER THREE

SYNTHESIZER DESCRIPTION

3.1 Introduction

In this chapter we shall discuss the properties of a terminal analog synthesizer and in particular the synthesizer used in this thesis work. Terminal analog synthesizers have been studied in great detail. Hardware synthesizers capable of producing high quality speech have been developed and are currently in use -- SPASS at MIT and OVE II in Stockholm are two examples. We have developed our own synthesizer which has been simulated on a large computer. A thorough explanation of the aspects of this synthesizer has been included in this chapter.

3.2 Terminal Analog Synthesizer Characteristics

A terminal analog synthesizer models the behavior of the vocal tract in terms of its overall transmission - as viewed from its input and output terminals. It must also generate appropriate sources of excitation, and account for the radiation characteristics at the output.

Let us first consider the transmission properties of the vocal tract.

For an unconstricted vocal tract (approximated as a uniform pipe) excited at the glottis, closed at the vocal cords and open at the mouth, the transmission ratio of mouth and glottal volume velocities has a frequency domain representation.

$$\frac{U_m}{U_g} = \frac{1}{\cosh \gamma(s) l} = \prod_{n=0}^{\infty} \frac{S_n S_n^*}{(S - S_n)(S - S_n^*)} \qquad (3.1)$$

where

Um = volume velocity at the mouth

Ug = volume velocity at the glottis

$$\gamma(s) = \left[(Ra + sLa)(Ga + sCa) \right]^{1/2}$$

AND

Ra, La, Ga, AND Ca

are the per-unit-length acoustical parameters of the vocal tract, and

ℓ = the length of the vocal tract.

The poles, for a straight pipe approximation to the vocal tract, are uniformly spaced along the $j\omega$ axis (assuming no losses). For a nonuniform pipe approximation, the transmission will generally have its poles spaced nonuniformly in frequency. A terminal analog models this situation by cascading individual, isolated electrical resonators which can be tuned to suitable frequencies.

When the source of excitation does not occur at the glottis, the transmission ratio of mouth and glottal volume velocities will exhibit poles and zeros. Thus for these sounds provision must be made for suitably tunable zeros in cascade with poles.

3.3 Glottal Source Characteristics

Flanagan has shown (1966, Chapter III) that the source for voiced sounds is approximately a high impedance, constant volume-velocity generator. The glottal source does not appreciably influence the vocal tract configuration. There is evidence, however, that the vocal tract configuration does influence the glottal source characteristics (Klatt, 1967). In most terminal analog synthesizers the glottal source characteristics are represented independently of the vocal tract

configuration. (The glottal source characteristics are fundamental frequency of vibration, pulse amplitude, and pulse shape.)

The properties of the glottal source wave are not well understood and generally they are approximated by choosing a wave shape whose amplitude spectrum falls at about 12 db/octave. Generally this is done by exciting a fixed, spectral-shaping network by repeated impulses.

3.4 Radiation Characteristics

To simulate the effects of the radiation transfer characteristic for the conversion from volume velocity at the mouth to pressure measured a fixed distance from the lips, a spectral shaping of +6 db/octave must be used.

3.5 Description of BLØDI Synthesizer

A block diagram of the synthesizer used throughout this thesis work is shown in Fig. 3.1. This synthesizer was simulated on an IBM 7094 digital computer using BLØDI (Block Diagram Compiler, Kelly and Vyssotsky, 1961).

The synthesizer has three parallel branches for processing signals. The upper branch is used for producing normal voiced speech and can also be used for whispered and/or aspirated speech. The upper branch consists of seven fixed resonators (F10 through F4), a shaping network $\left[\frac{s}{(s+a)(s+a^*)}\right]$, and five variable resonators (F1, F2, F3, NASAL POLE, NASAL ZERO). (The constant a used for the shaping network is a complex number.) The purpose of the shaping network is to produce both the correct pitch pulse shape and high frequency behavior of the speech spectrum. The characteristics of this network will be discussed later in this chapter.

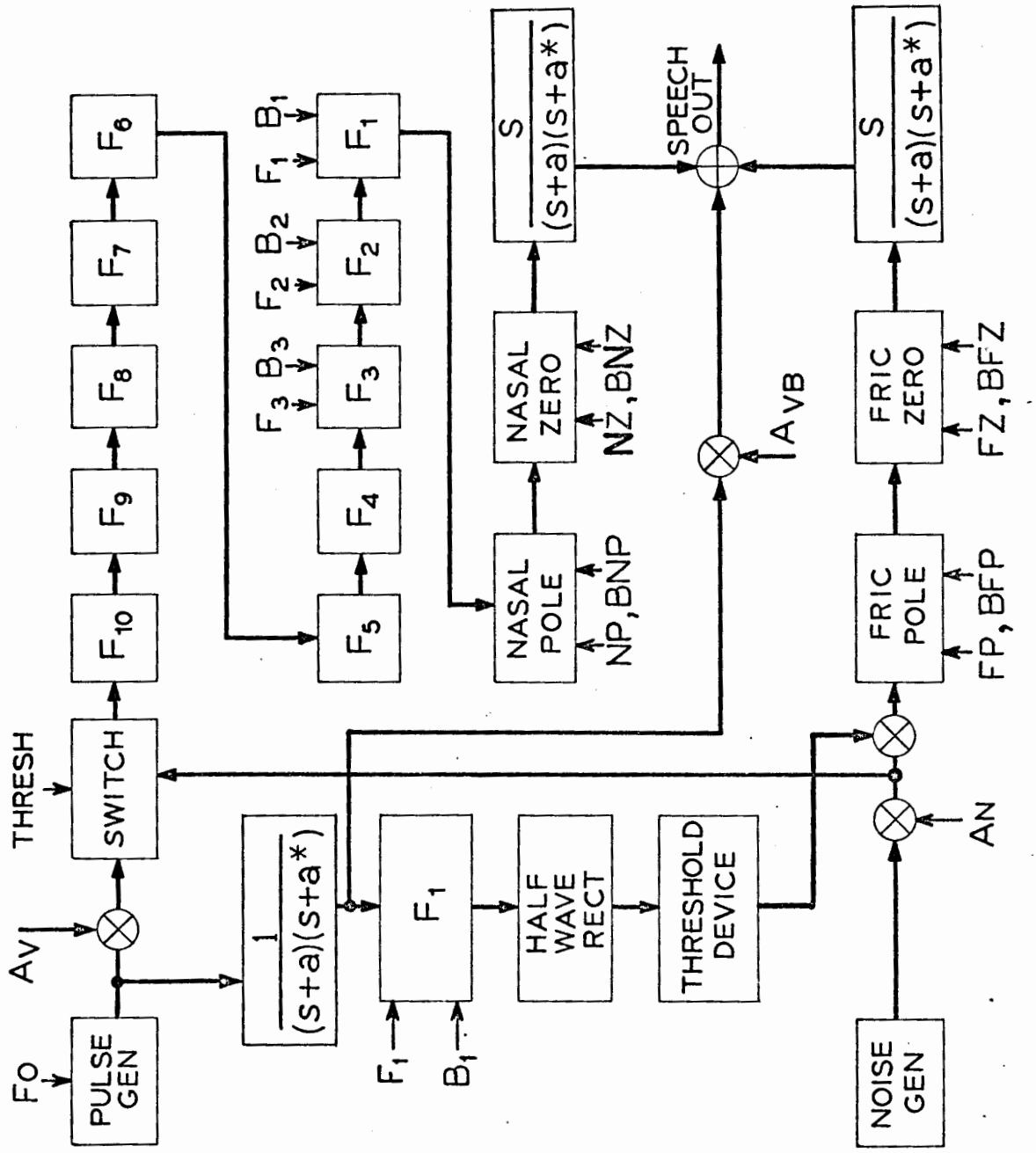


FIGURE 3.1

BLODI SIMULATED SPEECH SYNTHESIZER

The five variable resonators are externally controlled (both bandwidths and center frequencies of the resonators) and are used to represent three formants for voiced sounds and an additional pole-zero pair for nasals.

The raison d'être of the fixed resonators F4 to F10 will be explained in a later section.

The lower branch consists of two variable resonators (FRIC POLE and FRIC ZERO) and a shaping network $\left[\frac{s}{(s+a)(s+a^*)}\right]$ identical to the one described above. The variable resonators are externally controlled and are used in the production of fricatives and for the frication period of stop consonants.

The middle branch consists of another spectrum shaper $\left(\frac{1}{(s+a)(s+a^*)}\right)$, a variable resonator (F1) and a gain control (A_{VB}). This path is used to produce a low frequency voicebar for the quiet interval of voiced stop consonants.

Finally there is a cross path between the upper and lower branches of the synthesizer. This path is used for the production of the voiceless spectrum of the voiced fricatives |Z,ZH|. A detailed explanation of this process will be given later in this chapter.

There are two sources which can excite the synthesizer. These are a random white noise generator (NOISE GEN), and a single shot multivibrator (PULSE GEN) which puts out impulses (pitch pulses) at times specified externally. These two sources can be gated (with suitable weighting) to either the upper or lower branches.

The noise source is gated to the upper branch for aspiration for voiceless stop consonants, to produce an |H| sound, and to produce

whispered speech. The noise source is gated to the lower branch to produce the fricatives, and the frication period of stop consonants. The pulse generator is gated to the upper branch in the production of voiced speech and for the voiced spectrum of voiced fricatives. It is gated to the middle branch for the voicebar for stop consonants and gated to the lower branch for the voiceless part of the voiced fricatives |Z,ZH|.

The synthesizer control parameters are indicated in Fig. 3.1 by arrows. These include three amplitude controls, fourteen pole and zero controls (center frequencies and bandwidths), a fundamental frequency control, and a switch control. The amplitude controls (A_V, A_N, A_{VB}) set levels of voicing, noise, and voicebar. The voicebar is used during the stop gap of voiced stop consonants.

Six of the pole-zero controls are used for bandwidths and center frequencies of the three variable formant resonators (F_1, B_1, F_2, B_2, F_3, B_3). Four of the pole-zero controls are used to control center frequencies and bandwidths of the nasal pole and zero (NP, BNP, NZ, BNZ). When nonnasal sounds are produced, the cascade of nasal pole and nasal zero yields unity overall transmission. The remaining pole-zero parameters control center frequencies and bandwidths of the fricative pole and zero (FP, BFP, FZ, BFZ).

The fundamental frequency control (FO) sets times at which impulses excite the voicing network. The switch control (THRESH) determines whether the upper branch of the synthesizer is excited by noise or impulses.

The levels of the sources (NOISE GEN and PULSE GEN) are set so that they supply approximately equal power. Since the average power in voiced speech is about four times the average power in unvoiced speech, the noise amplitude control (A_N) for synthesizing a fricative consonant would be set to about half the value of voicing amplitude (A_V) for synthesizing a vowel in a typical case.

The speech out is a direct sum of the outputs from the three branches. The synthesizer produces output samples every 1/20,000th of a second. Therefore, the variable resonators can have resonances up to 10 kHz.

3.6 Features of the Simulation

The synthesizer was simulated on a computer. There are two basic building blocks and these are the formant resonator (FMT) and a box which produces a pair of complex conjugate zeros (ZER). We will discuss these boxes in detail.

3.6.1 Formant Resonator - FMT

The FMT box produces a pair of complex conjugate poles. The analog transfer function of this box is:

$$H(s) = \frac{s_i s_i^*}{(s+s_i)(s+s_i^*)} \qquad (3.2)$$

where

$$s_i = \sigma_i + j\omega_i$$
$$s_i^* = \sigma_i - j\omega_i$$

We see that the DC gain of H(s) is unity so we can cascade a number of these boxes and still have a DC gain of unity. (The DC gain of the vocal tract is unity and we are modelling this effect here.)

A simple electrical network for realizing the transfer function of Eq. (3.2) is seen in Fig. 3.2. From Fig. 3.2 we get

$$\frac{e_o}{e_i} = \frac{\frac{1}{Cs}}{\frac{1}{Cs} + Ls + R} = \frac{\frac{1}{LC}}{s^2 + \frac{R}{L}s + \frac{1}{LC}} \qquad (3.3)$$

$$\omega_i \approx \frac{1}{\sqrt{LC}} \quad , \quad \sigma_i = \frac{R}{2L} \qquad (3.4)$$

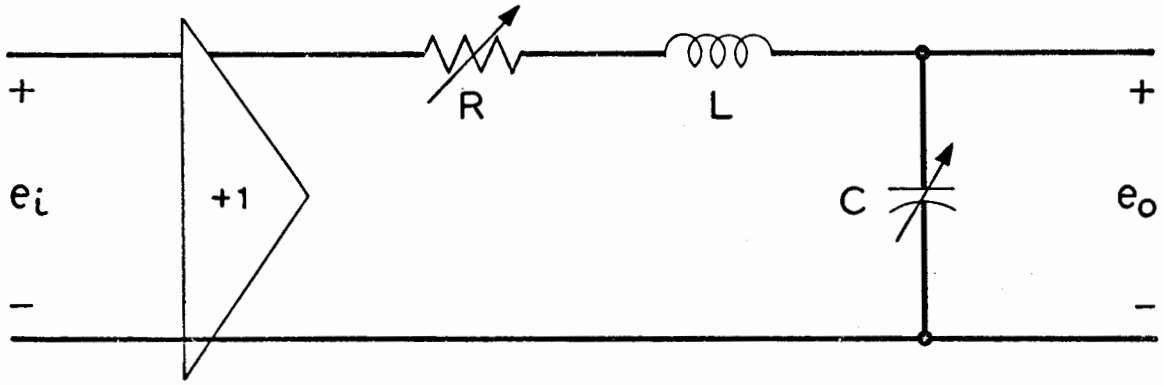
In order to obtain a simulated resonator, we must apply the z transform to the transfer function of Eq. (3.2). Equation (3.2) can be rewritten as

$$H(s) = \frac{\sigma_i^2 + \omega_i^2}{s^2 + 2\sigma_i s + \omega_i^2 + \sigma_i^2} \qquad (3.5)$$

Applying the impulse invariant z transform to H(s) (see Radar and Gold, 1966) we get

$$H(z) = \frac{\left(\frac{\sigma_i^2 + \omega_i^2}{\omega_i}\right) e^{-\sigma_i T} \sin(\omega_i T) z^{-1}}{1 - 2e^{-\sigma_i T}\cos(\omega_i T) z^{-1} + e^{-2\sigma_i T} z^{-2}} \qquad (3.6)$$

where $T = \frac{1}{F_s} = \frac{1}{20000}$ sec = sampling period.



$$\frac{e_o(s)}{e_i(s)} = \frac{\frac{1}{LC}}{s^2 + \frac{R}{L}s + \frac{1}{LC}}$$

$$\frac{e_o(s)}{e_i(s)} \cong \frac{\omega_i^2}{s^2 + 2\sigma_i s + \omega_i^2} \qquad (\sigma_i << \omega_i)$$

$$\omega_i = \frac{1}{\sqrt{LC}} \qquad \sigma_i = +\frac{R}{2L}$$

FIG. 3.2 ELECTRICAL RESONATOR

If we consider Eq. (3.6) as a transfer function between $e_i(nT)$ and $e_o(nT)$ we obtain the difference equation

$$e_o(nT) = 2e^{-\sigma_i T} \cos \omega_i T \, e_o(nT-T) - e^{-2\sigma_i T} e_o(nT-2T) + \left(\frac{\sigma_i^2 + \omega_i^2}{\omega_i}\right) e^{-\sigma_i T} \sin \omega_i T \; e_i(nT-T) \qquad (3.7)$$

A simple realization of this difference equation is seen in Fig. 3.3.

The resonators used in this simulation were not identical to those of Fig. 3.3. The reason for this can be seen from an examination of Eq. (3.6). We have stated that the vocal tract transmission function has a DC gain of unity and hence we would like to cascade resonators with a unity DC gain. The digital resonator of Eq. (3.6) has a DC gain ($z^{-1} = 1$) of

$$H(1) = \left(\frac{\sigma_i^2 + \omega_i^2}{\omega_i}\right) \frac{e^{-\sigma_i T} \sin \omega_i T}{1 - 2e^{-\sigma_i T} \cos \omega_i T + e^{-2\sigma_i T}} \qquad (3.8)$$

The gain is clearly a function of the pole position (σ_i, ω_i) . To alleviate this problem, an additional multiplicative factor was applied to give the desired unity gain at DC.

3.6.2 Complex Conjugate Zero Pair-ZER

The purpose of the zero box is to provide a pair of complex conjugate zeros. The analog transfer function is

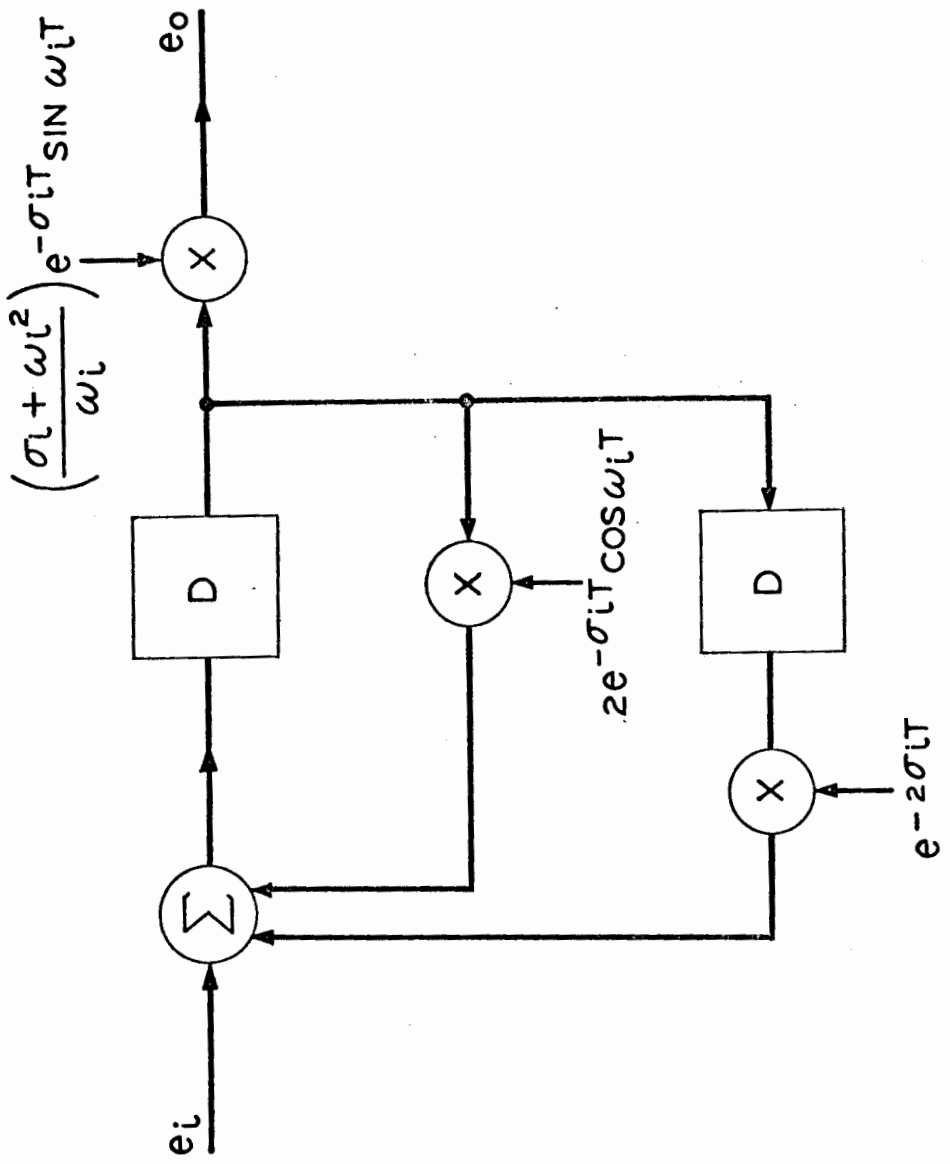


FIG. 3.3 DIGITAL FILTER REALIZATION OF ELECTRICAL RESONATOR

$$H_1(s) = \frac{(s-s_1)(s-s_1^*)}{s_1 s_1^*} = \frac{s^2 + 2\sigma_i s + \sigma_i^2 + \omega_i^2}{\sigma_i^2 + \omega_i^2} \qquad (3.9)$$

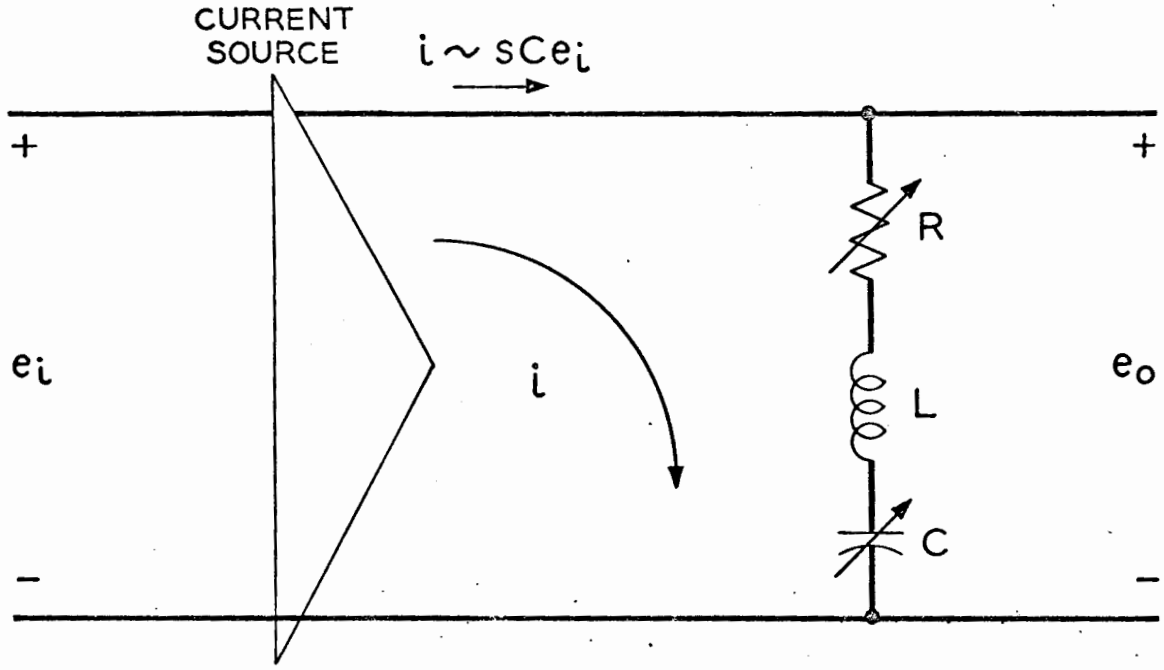
Again we see that the transfer function has unity DC gain.

An electrical network to realize this transfer function is seen in Fig. 3.4. From Fig. 3.4 we get

$$\frac{e_o}{e_i} = sC\left[R + Ls + \frac{1}{Cs}\right] = \frac{s^2 + s\frac{R}{L} + \frac{1}{LC}}{\frac{1}{LC}} \qquad (3.10)$$

The computer simulated version of this filter can be obtained as the inverse of the FMT filter or

$$H_1(z) = \left(\frac{\omega_i}{\sigma_i^2 + \omega_i^2}\right)\left[\frac{1 - 2e^{-\sigma_i T}\cos\omega_i T\, z^{-1} + e^{-2\sigma_i T} z^{-2}}{z^{-1} e^{-\sigma_i T} \sin\omega_i T}\right] \qquad (3.11)$$



$$\frac{e_o(s)}{e_i(s)} = \frac{s^2 + \frac{sR}{L} + \frac{1}{LC}}{\frac{1}{LC}}$$

$$= \frac{s^2 + 2\sigma_i s + \omega_i^2}{\omega_i^2} \qquad (\sigma_i << \omega_i)$$

$$\omega_i = \frac{1}{\sqrt{LC}} \quad ; \quad \sigma_i = \frac{R}{2L}$$

FIG. 3.4 ELECTRICAL CIRCUIT TO PRODUCE A CONJUGATE PAIR OF ZEROS

The z^{-1} factor in the denominator makes this transfer function unrealizable as it implies the output arrives one sample before the input, so we must realize a new transfer function

$$H_2(z) = z^{-1} H_1(z) = \frac{e_0(z)}{e_i(z)} \tag{3.12}$$

The difference equation corresponding to Eq. (3.12) is

$$e_0(nT)\, e^{-\sigma_i T} \sin \omega_i T \left(\frac{\sigma_i^2 + \omega_i^2}{\omega_i}\right) = e_i(nT)$$

$$-2e^{-\sigma_i T} \cos \omega_i T \; e_i(nT-T) + e^{-2\sigma_i T} e_i(nT-2T) \tag{3.13}$$

The block diagram realization of this equation is seen in Fig. 3.5.

Again the multiplicative constant was changed so as to give the digital filter a unity DC gain.

The cascade of a zero box and a pole box - with identical multiplicative constants, will yield an overall transmission of unity.

A zero box must not be placed in the system too near the sources of excitation. This is because a zero box is not low pass in character. It has a high frequency boost of 12 db/octave. In sampled data systems energy present at frequencies above Fs/2 is aliased down to a frequency

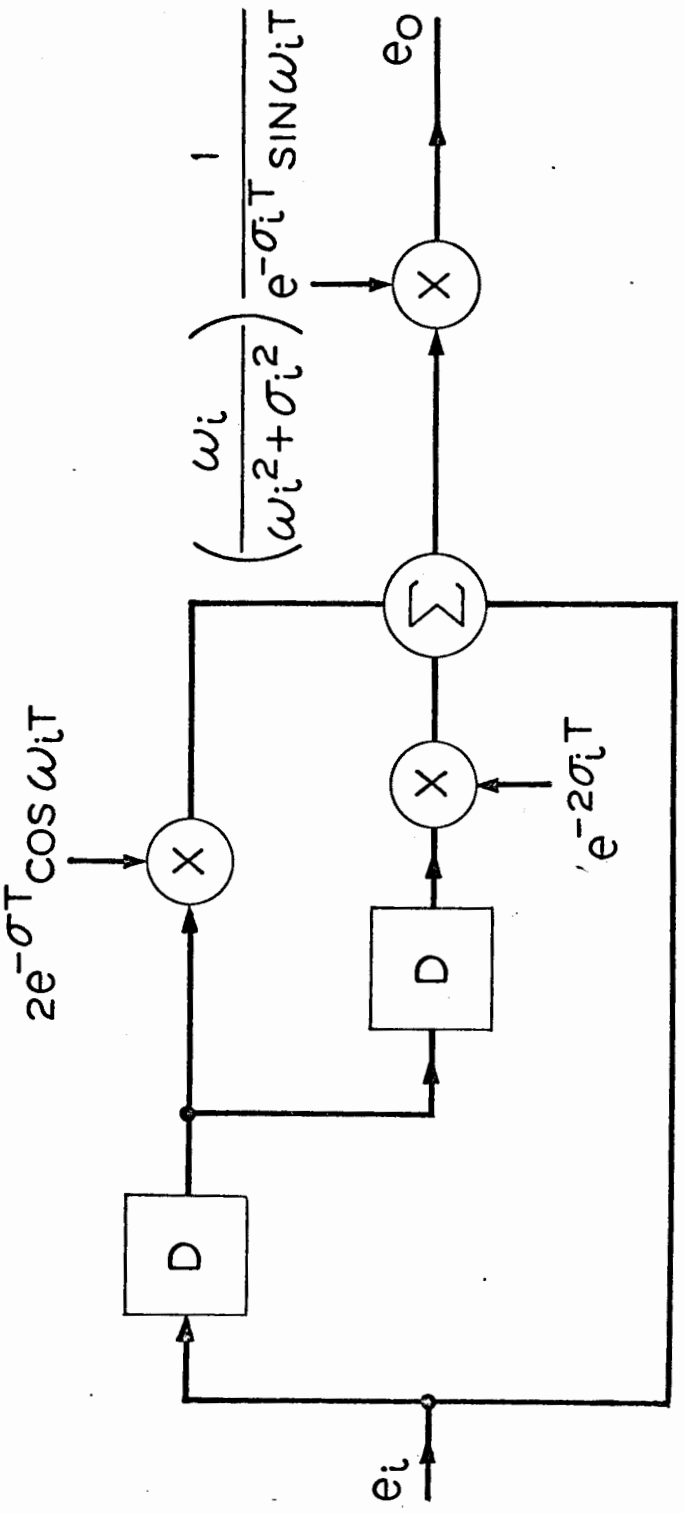


FIG. 3.5 DIGITAL FILTER REALIZATION OF A CIRCUIT TO PRODUCE A PAIR OF COMPLEX CONJUGATE ZEROS

(Fs-Fs/2). Any time limited waveform must have some energy above Fs/2. Generally the energy above Fs/2 is much lower in level than the energy below Fs/2 (because of suitable filtering before sampling). However if the sampled waveform is processed by elements which are not low pass in character (i.e., a zero box) then the energy above Fs/2 is boosted in level and will make the aliasing problems more severe.

3.7 Individual Aspects of The BLØDI Synthesizer

There are three further aspects of the synthesizer that bear discussion. These are (1) the resonators F4 through F10, (2) the cross connecting arm of the synthesizer and (3) the shaping filter.

3.7.1 F4 through F10

The transmission function of the vocal tract of volume velocity from glottis to mouth has an infinite set of singularities. In other words there are an infinite set of resonances of the vocal tract. A terminal analog synthesizer, of necessity, can have only a finite number of resonances. Most analog synthesizers use four or five resonances in cascade with a correction network referred to as a higher-pole correction. The higher-pole correction network is necessary to account for the effects of higher order poles in the low frequency region of interest (usually in the region 0-5 kHz). The characteristics of this network are derived from the assumption that the higher order resonances that were neglected are at the fixed frequencies appropriate to an unconstricted vocal tract, i.e., odd multiples of 500 Hz. Flanagan, (1965), under these assumptions, has derived asymptotic approximation to the frequency behavior of this higher pole correction. His final result is that

$$\left|Q_K(j\omega)\right| \approx e^{(\omega/\omega_1)^2}\left[\frac{\pi^2}{8} - \sum_1^K \frac{1}{(2n-1)^2}\right] \qquad (3.14)$$

where K is the number of reasonances actually used and ω_1 is the lowest resonance of the unconstricted vocal tract ($2\pi\times500$ rad/sec). Circuits have been built to approximate the frequency characteristics of Eq. (3.14) for frequencies up to about 4.5-5 kHz. Above this frequency, people are not interested, generally, in faithfully reproducing the transmission properties of the vocal tract.

The need for a higher pole correction did not arise in the synthesizer described above. This is the reason for resonators F4 through F10. The resonance frequencies, bandwidths and Q's of these filters are tabulated in Table 3.1. (Values were extrapolated from a graph by Dunn, 1961.) We see that the center frequencies of the filters are at odd multiples of 500 Hz from 3500 Hz to 9500 Hz. It would appear that all we have done is included the first ten resonances and we are still lacking the infinite set of resonances beginning at 10500 Hz. This is not quite the case. We are dealing with a sampled data system which is both periodic in frequency (period = sampling frequency = 20000 Hz) and symmetric about half the sampling frequency or 10000 Hz. (i.e., a pole at frequency F1 has a mirror image pole at a frequency 20000-F1.) Thus the resonance for F10 at 9500 Hz has periodicity poles at 29500, 49500, 69500, ... Hz and mirror image poles at 10500, 30500, 50500, ... Hz.

TABLE 3.1

RESONATOR FREQUENCIES AND BANDWIDTHS

RESONATOR	CF (Hz)	Q	BW (Hz)
F_4	3500	20	175
F_5	4500	16	281
F_6	5500	12	458
F_7	6500	9	722
F_8	7500	6	1250
F_9	8500	4	2125
F_{10}	9500	2	4750

(EXTRAPOLATED FROM DUNN- 1961)

Similarly all other resonators have both periodicity and mirror image poles and thus we have an infinite set of poles which are approximately at odd multiples of 500 Hz. Thus the need for a correction term has been eliminated.

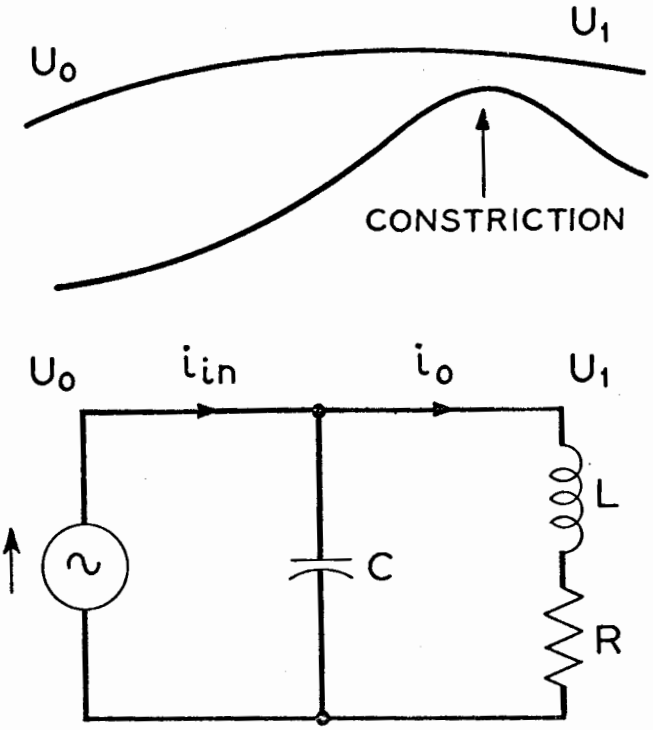
3.7.2 Cross Connecting Arm of The Synthesizer

We discussed in Fig. 3.1 the existence of a path connecting the pulse generator source to the noise source by means of suitable networks. This is used to synthesize voiced fricatives.

A voiced fricative is characterized by a low intensity voiced formant structure and a characteristic noise spectrum which is amplitude modulated by a suitably shaped pitch pulse. To model this type of effect, most previous synthesizers have merely added a voicing component to an unvoiced component. The dissatisfaction with the results attained led to a reexamination of the method of production of voiced fricatives.

A simplified explanation of how voiced fricatives are produced is as follows. The vocal cords are vibrating, sending out puffs of air (pitch pulses) at given times. These puffs of air move through a back cavity from the glottis to a point of constriction in the vocal tract. As air passes through the constriction, turbulence is produced. The turbulent air then passes through the cavity in front of the constriction and is released through the mouth. The lower frequency voiced spectrum comes from the nonturbulent component of air which also passes through the front cavity and out of the mouth.

Figure 3.6 shows a first order representation for the volume velocity transfer function and the area function for voiced fricatives. Figure 3.6a shows a schematized articulatory structure for producing voiced fricatives. U_o is the volume velocity at the glottis (represented



$$\frac{i_o}{i_{in}} = \frac{U_1}{U_0} = \frac{1}{LCs^2 + RCs + 1} \cong \frac{F_1^2 (2\pi)^2}{s^2 + 2\pi B_1 s + F_1^2 (2\pi)^2}$$

F_1 = FORMANT ONE CENTER FREQUENCY (HZ)

B_1 = BANDWIDTH OF FORMANT ONE (HZ)

FIGURE 3.6

(a) SCHEMATIZED ARTICULATORY STRUCTURE FOR PRODUCTION OF VOICED FRICATIVES

(b) ELECTRICAL EQUIVALENT CIRCUIT FOR CALCULATING VOLUME VELOCITY TRANSFER FUNCTION

at the far left in Fig. 3.6a) and U_1 is the volume velocity at a point just beyond the constriction. Between the glottis and the constriction lies a back cavity. An electrical equivalent circuit for calculating the volume velocity transfer function of Fig. 3.6a is shown in Fig. 3.6b. The back cavity is approximated by a capacitor (C); losses are represented by the resistor (R); and the constriction is approximated by the inductor (L). A single complex conjugate pole pair transfer function is attained. The pole position is identical to the position of the first formant of the voiced spectrum. A more complete representation would take into account the higher frequency poles of the transfer function. This is unnecessary because the first pole is at a low frequency (characteristic of a highly constricted vocal tract), and the amplitudes of the higher frequency poles are small due to the 12 db per octave attenuation from the lowest pole. Furthermore the glottal source spectrum for voiced fricatives provides additional high frequency attenuation.

These physical effects are modelled in the synthesizer by the connection between the pitch pulse generator and the lower branch. (Refer again to Fig. 3.1.) The pitch impulse is passed through a shaping filter to produce a pitch pulse. (This will be discussed in the next section.) [The cutoff frequency of the shaping filter is identical to that used in normal vowel production. This is probably incorrect since a constricted vocal tract would interact more strongly with the glottal source than a unconstricted one. Lack of a proper rationale for choosing a different cutoff frequency necessitated leaving it the same.] The pitch pulse is then passed through a resonator which is externally tuned to formant one, the actual first formant of the voiced spectrum. In order to produce

turbulence physically, the volume velocity of the air flow must exceed some critical value. In an effort to model this type of threshold effect, the threshold device and half-wave rectifier have been used. The output of the threshold device is a constant whenever the input is less than some critical value. Whenever the input exceeds this critical value, the output equals the input. The final step in the process is a multiplication of the waveform at the output of the threshold device by noise. The final output can be considered as noise nonlinearly modulated by a formant one shaped pulse which is synchronous with the pitch pulse. All indications are that the voiced fricatives are of high quality using this method. (See also Maxey, 1963, for a further discussion of a method for production of voiced fricatives on a terminal analog synthesizer.)

3.7.3 Pitch Pulse Shaping

For voiced sounds, the vocal tract is excited by a stream of pulses. In our synthesizer we produce these pulses by exciting a shaping filter with impulses. The shape of the pitch pulses can be seen in Fig. 3.7. Line 1 in Fig. 3.7 shows the input impulses and line 2 shows the pitch pulses so attained. We see that there is overshoot in these pulses. This is because the impulses were put through a filter of the form

$$H(S) = \frac{1}{(S+\sigma a + j\omega a)(S+\sigma a - j\omega a)} \qquad (3.15)$$

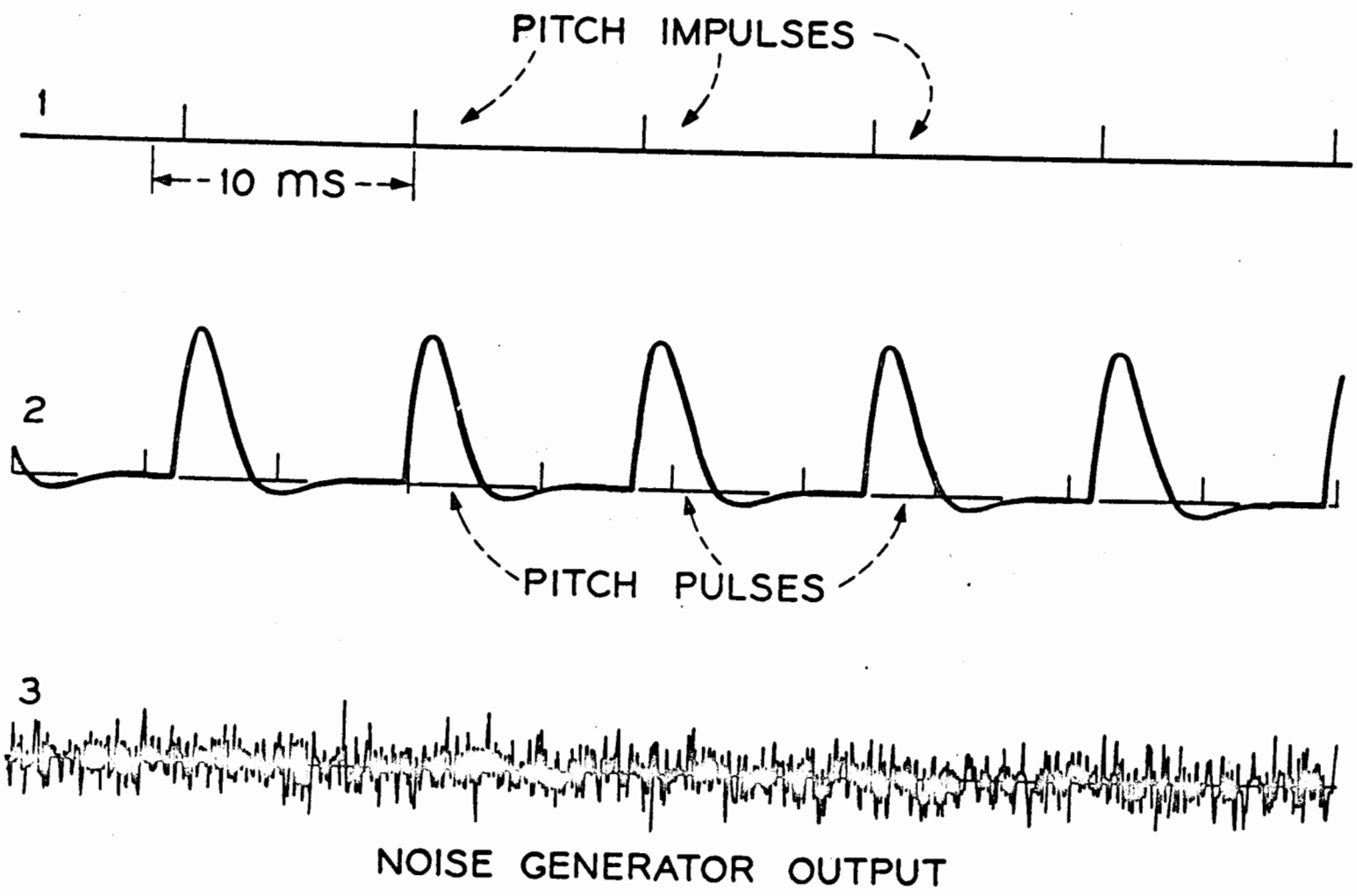


FIGURE 3.7

(1) TIMES OF OCCURRENCE OF IMPULSES WHICH EXCITE PITCH PULSE SHAPING FILTER

(2) PITCH PULSE SHAPES – REPRESENTATION OF EQUATION 3.16

(3) RANDOM NOISE GENERATOR OUTPUT

where

$$\sigma a = 250\pi \text{ RAD/SEC}$$
$$\omega a = 400\pi \text{ RAD/SEC}$$

The impulse response of H(s) is

$$h_a(t) = \frac{e^{-\sigma a t}}{\sigma a} \text{ SIN } \omega a\, t, t \geq 0 \qquad (3.16)$$

Line 2 of Fig. 3.7 shows pulses of the form of Eq. (3.16). (Line 3 of Fig. 3.7 is a typical example of the output of the noise generator of the synthesizer of Fig. 3.1.)

The shaping filters of Fig. 3.1 are of the form $\frac{1}{(s+a)(s+a^*)}$ where a is a complex number. This complex conjugate pole pair was used instead of a real axis double order pole as a result of intensive investigations by C. Coker at Bell Telephone Laboratories (personal communication) on the proper placement of this pole. He found that a more natural sounding voice was attained with the poles off the real axis. His values for real and imaginary parts were accepted and used for this synthesizer. The shape of the pitch pulse corresponds closely to pitch pulse shapes observed from other techniques.

3.8 Simulation Versus Hardware

The synthesizer described above was computer simulated. Appendix A contains a BLØDI listing of the synthesizer with some explanation as to terminology.

There is always a debate as to whether the synthesizer should be built in hardware or designed entirely for computer simulation. For this research project the advantages of computer simulation were well appreciated. The formant of the synthesizer (its block diagram configuration) has been changing periodically since this thesis research began. As new ideas were generated, new synthesizer configurations had to be tested. The ease of change and flexibility of a computer program are strong reasons for computer simulation. The entire synthesizer could be changed by merely inserting, removing or changing cards in a deck of computer cards. With a hardware synthesizer, it might take weeks to make such changes and should the changes prove undesirable it would take weeks to return to its old configuration.

Another advantage of computer simulation is that one can monitor waveforms at any point in the synthesizer and retain a permanent copy.

The major disadvantage to computer simulation is the long run time required to produce the speech. The present version of the synthesizer takes about 50-60 times real time to generate the speech. When computer time is an important factor, this must be taken into consideration.

3.9 Parallel Versus Serial

As mentioned previously, one can use the synthesizer as a cascade of resonant circuits (poles and zeros) or a parallel combination

of resonant circuits. The parallel combination is a means of matching a generated spectrum with a desired one. It does not parallel the speech producing mechanism from a source-system point of view. The source of excitation for a parallel synthesizer must have a flat spectrum. Hence impulses are used as a source. This is not a faithful representation of the glottal source for voiced sounds. If one wanted to control the glottal pulse shape (or spectrum) for synthesis by rule, then the configuration of a parallel synthesizer would have to undergo serious modification. Furhtermore, the transfer function for vowels with a parallel synthesizer contains zeros as well as poles. A real vowel transfer function contains only poles. The presence of undesired zeros tends to limit the applicability of parallel synthesizers.

There are advantages of parallel synthesizers. Since there is independent control of formant amplitudes, the constant spectra can be accurately reproduced. (However, the additional amplitude controls make the rules for synthesis somewhat more complicated.) There is no need for a higher pole correction for parallel synthesizers as the high frequency behavior has the correct form without a correction network. Furthermore, noise propagates additively in a parallel synthesizer, whereas it propagates multiplicatively in a serial synthesizer. Hence sudden changes in formant center frequencies and/or bandwidths produce less objectionable audible effects for a parallel synthesizer than for comparable changes in a serial synthesizer.

The serial synthesizer has been used in this thesis. The natural ways in which source parameters can be controlled and the elimination of the need for additional ad hoc rules for determining

formant amplitudes were among the reasons for our choice. There are advantages and disadvantages of both types of synthesizers.

3.10 Preliminary Test of the Synthesizer

The adequacy of the synthesizer was examined informally. The synthesizer was tested with data extracted from a formant analyzer. (Only fundamental frequency, formant frequencies, amplitudes of noise and voicing, and fricative pole and zero positions were extracted by the formant analyzer.) The sentence used was, "I am speaking dumb and vain words." In the opinion of the author and an engineer with experience in speech work, the synthesized speech was of good quality. The synthesizer was tested further with an utterance where certain parameters were extracted by hand from a spectrogram. The utterance was, "You are a real liar." The opinion of the author and the engineer was that the speech was again of good quality. These tests were intended to show that the synthesizer could produce intelligible speech and are not a thorough test of the synthesizer. [Examples of these utterances are available on a demonstration tape which accompanies this thesis.] The adequacy of terminal analog synthesizers has been demonstrated independently by other researchers (e.g., Fant, Holmes).

3.11 Summary of Control Signals

In summary the synthesizer control signals are as follows:

		Symbol Used In Thesis
1.	Pitch Period or Fundamental Frequency	TO or FO
2.	Amplitude of Voicing	A_V
3.	Amplitude of Noise	A_N
4.	Formant 1--Center Frequency	F_1
5.	Formant 2--Center Frequency	F_2
6.	Formant 3--Center Frequency	F_3
7.	Nasal Pole--Center Frequency	NP
8.	Nasal Zero--Center Frequency	NZ
9.	Fricative Pole--Center Frequency	FP
10.	Fricative Zero--Center Frequency	FZ
11.	Formant 1--Bandwidth	B_1
12.	Formant 2--Bandwidth	B_2
13.	Formant 3--Bandwidth	B_3
14.	Nasal Pole--Bandwidth	BNP
15.	Nasal Zero--Bandwidth	BNZ
16.	Fricative Pole--Bandwidth	BFP
17.	Fricative Zero--Bandwidth	BFZ
18.	Decision Control for Gate	THRESH
19.	Amplitude of Voicebar	A_{VB}

Control signal 18 is a binary signal. If its level is 100 (the normal level) the pitch pulse goes through the upper path of the synthesizer. If its level is 0, the noise goes through the upper path - as in whispered speech.

CHAPTER FOUR

QUASI-STATIC REPRESENTATIONS OF PHONEMES AS APPLIED TO THE SYNTHESIS STRATEGY

4.1 Introduction

A major purpose of this thesis is to provide a method of converting from linguistic symbols to control signals as a function of time. Chapter five contains the discussion of the scheme we have adopted for this purpose. This chapter provides the bridge between the synthesizer of Chapter three and the synthesis scheme.

In this chapter we shall discuss the classes of phonemes and their synthesis characteristics. The discussion will lead to quasi-static representations of the phonemes. These input representations will not always be reflections of what will occur during the dynamics of speech production. For instance consonants are represented by target values of three formants. These formant resonances need never be excited during the production of the consonant. Thus the actual phoneme realization need not be identical to the input representation.

The characterizations of the phonemes developed in this chapter will be used in our synthesis strategy.

4.2 Vowels

The theory of production of vowels is straightforward. The source is at the glottis and the vocal tract acts like a pipe with varying cross section area. An all-pole transfer function for vowels is thus attained. The pole positions for any given vowel are a function of sex, age, vocal tract characteristics, and context of the utterance in which the vowel occurs (Peterson, Barney, 1952, Stevens, 1966).

A single speaker, when producing the same vowel in different contexts, will have different pole positions depending on context. For example, Stevens (1966) has shown that pole positions for the I in B-I-L and the I in isolation are different.

The notion that the same vowel can be characterized by different sets of formants is non-trivial. Kim (1966) has worked on a technique for generating new sets of formant frequencies for vowels as a function of such factors as sex, and age. Basically he generates vowel formant frequencies from phonetic (articulatory) information about the vowel (high, mid, low, front, back, etc.) and a center of gravity for each formant. To change his vowel formants he merely has to change his center of gravity and the weighting factor for each of his phonetic attributes.

In our synthesis strategy vowels are characterized by sets of target values of three formant frequencies. These target values represent average sets of adult male vowel formant frequencies. Steady state vowels synthesized from these formant values should be highly identifiable. (Steady state means no synthesizer parameters vary with time.) The effects of context on vowel formant frequencies will be included in the dynamic aspects of our synthesis strategy.

The possible vowel inputs we have considered are IY, I, E, AE, A, UH, U, OW, OO, ER. According to Lehiste and Peterson (1961), the vowels I, E, U, and UH are called lax vowels. A lax vowel is generally characterized by a short transition from the preceding phoneme. The vowels IY, AE, A, OW, OO, and ER are called tense vowels. These are characterized by a relatively long transition from the preceding sound. The pairs of vowels (IY, I), (OO,U) and (A, UH) are phonetically similar in all ways except that the first of each pair is tense and the second lax.

Table 4.1 gives a list of the vowel formant frequencies used in our synthesis scheme. The bandwidths of formants one, two, and three are 60, 100, and 120 Hz respectively.

The duration of a stressed vowel is primarily a function of the following consonant. House (1961) made a quantitative study of the duration of stressed vowels as a function of the following consonant. He showed that the duration of a vowel followed a well organized stratagem. The longest vowels were followed by voiced fricatives whereas the shortest vowels were followed by voiceless stops. A summary of House's data is presented in Table 4.2. (Vowel durations are in milliseconds.)

4.3 Diphthongs

There is general ambiguity and disagreement as to what is and what is not a diphthong. According to Webster's Seventh New Collegiate Dictionary a diphthong is "a gliding monosyllabic speech item that starts at or near the articulatory position for one vowel and moves to or toward the position for another." According to this definition there are five diphthongs in American English. These are e^{I}, o^{U}, aI, aU, and ɔI. According to many phoneticians this is not the case. They classify the sounds e^{I} and o^{U} as tense vowels. The exact status of e^{I} and o^{U} is unclear. For purposes of this thesis, they are treated as diphthongs. Some acoustical justification for this choice comes from the experiment described below.

Holbrook and Fairbanks (1962) made an analysis of data of the sounds e^{I}, o^{U}, aI, ɔI, and aU. They measured formant frequencies at different times during these sounds. Figure 4.1 shows their findings. The arrows in this figure indicate the direction (in the F1, F2 plane)

PHONEME	F1	F2	F3
IY	270	2290	3010
I	390	1990	2550
E	530	1840	2480
AE	660	1720	2410
UH	520	1190	2390
A	730	1090	2440
OW	570	840	2410
U	440	1020	2240
OO	300	870	2240
ER	490	1350	1690
W	300	610	2200
L	380	880	2575
R	420	1300	1600
RO	295	845	1315
Y	300	2200	3065

TABLE 4.1

FORMANT FREQUENCIES FOR VOWELS AND W, L, R, Y

(VOWEL DATA FROM PETERSON, BARNEY-1952) (OTHER DATA IN TABLE FROM LEHISTE, PETERSON-1962)

FOLLOWING CONSONANT	VOWEL									
	IY	I	E	AE	UH	A	OW	U	OO	ER
VOICED	315	215	250	375	250	365	355	220	320	340
UNVOICED	145	120	165	225	140	220	215	125	160	180
FRICATIVE	260	185	240	340	225	340	310	200	270	265
STOP	200	145	175	270	165	255	265	145	200	240
AVERAGE DURATION	225	160	200	300	200	240	280	170	240	250

TABLE 4.2

VOWEL DURATIONS AS A FUNCTION OF FOLLOWING CONSONANT

(FROM HOUSE-1961)

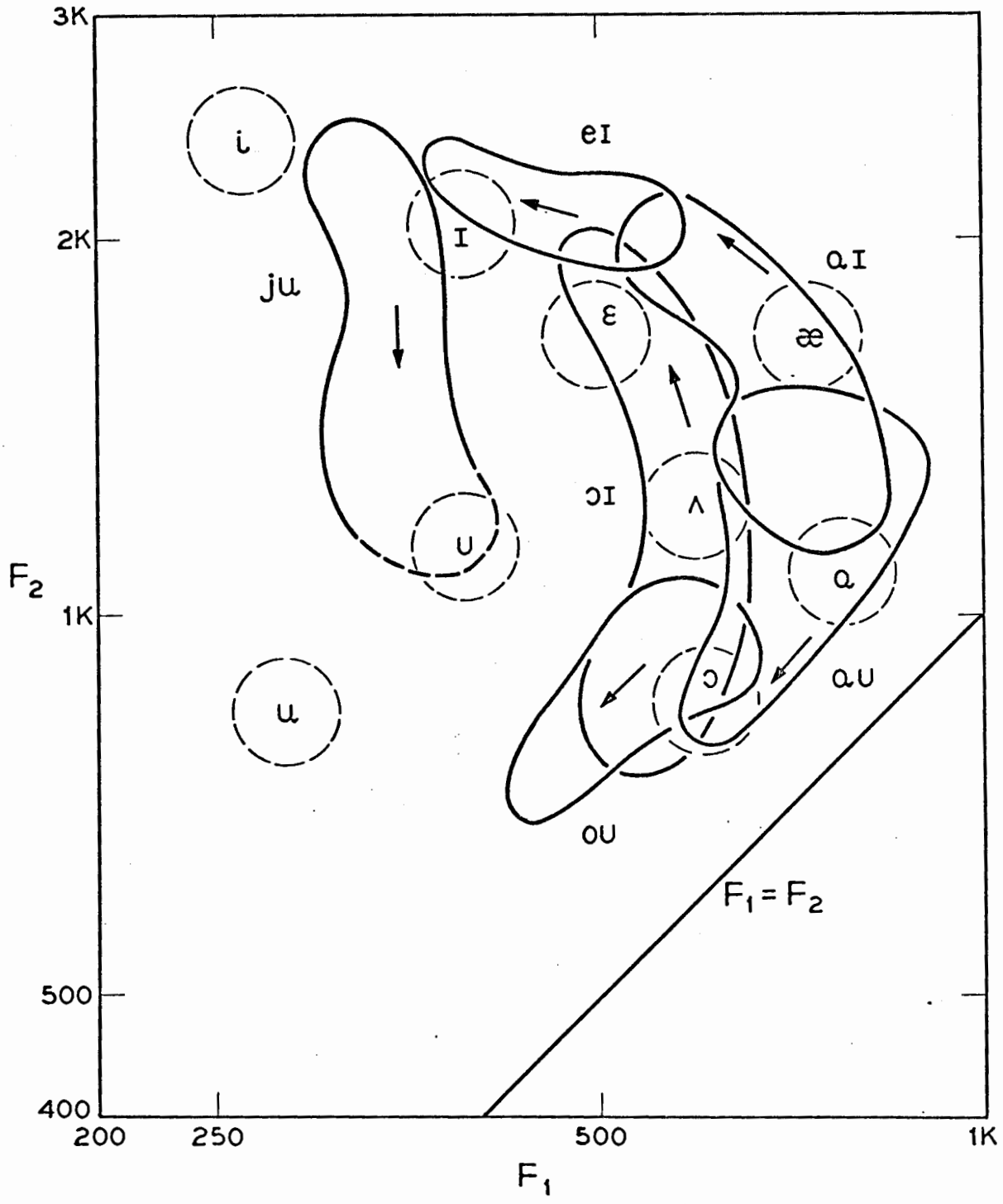


FIGURE 4.1

CENTRAL REGIONS OF VARIATION OF FREQUENCIES OF FORMANTS ONE AND TWO IN DIPHTHONGS (FROM HOLBROOK, AND FAIRBANKS-1962)

of motion of the formants. (The dashed circles in Fig. 4.1 indicate the positions of the vowels.) The data of Fig. 4.1 indicates that all five of these sounds can be represented as two target sequences. Holbrook and Fairbanks point out that for e^{I} and o^{U}, the duration of the initial target steady state is short.

The sounds e^{I} and o^{U} have been treated as sequences of E → IY, and OW → OO. (The choice of vowels in these sequences was determined from experiments with synthetic speech by the author.) The first vowel of these pairs has been drastically shortened so the first target is actually a virtual one in that no steady state formants exist until the second target is reached (if it is ever reached).

The sounds aI, ɔI, and aU have been treated here as two vowel sequences. The sequences used are: A → IY for aI, OW → IY for ɔI, and A → OO for aU. It is the second vowel of these pairs that has been shortened so whenever these diphthongs are increased in duration, as by stress, the increase in duration is manifested in the region of the first vowel target. Furthermore the transition time between the first and second vowels is increased as the duration of the diphthong increases. Finally formant 3 of the second vowel is restricted to lie close to formant 3 (within 200 Hz) of the first vowel. This restriction was necessitated to make the two vowel sequence better approximate the true diphthong formants without inventing a new set of vowels for this purpose.

4.4 W, L, R, Y

The group of sounds consisting of W, L, R, and Y is quite difficult to characterize. W and Y are sometimes called semivowels indicating their vowel-like character. They are generally characterized by short to moderate steady states and by slow gliding transitions to

adjacent phonemes. To all intents and purposes W and Y can be treated similar to the vowels with respect to the synthesis strategy. They are characterized by target positions of three formants.

R and L present different synthesis problems. The formant positions associated with these sounds are strongly a function of context. Lehiste and Peterson (1962) have studied formant positions for R and L as a function of word position. They found at least four allophones of L (initial position, intervocalic, final position, and syllabic) each having different formant structures. The case for R was significantly worse where they found twelve distinct allophones.

For our synthesis strategy R and L are represented by average values of three formant target positions. An initial allophone of R, referred to as RO, is also included in the representation of R. All other allophonic variations of formant positions for R and L have been neglected. The formant target values for W, L, R, RO, and Y are included in Table 4.1.

4.5 Nasals

The nasal consonants M, N, and NG are produced with the source of excitation at the glottis (the vocal cords are vibrating) and the vocal tract totally constricted at some point along the oral passageway. The velum is lowered so the air stream is released from the nose after traversing the nasal cavities. The totally constricted oral cavity acts as a side branch resonator. This side branch resonator causes the transfer function of the vocal tract for nasals to exhibit zeros, as well as poles. Furthermore, nasal consonants and nasalized vowels are characterized by resonances which are spectrally broader, or more highly damped, than

those for vowels. The broadening of the resonances is due to the nasal cavity through which the air must flow. The inner surface of the nasal tract is convoluted so that the nasal cavity has a relatively large ratio of surface area to cross-sectional area. Viscous and heat conduction losses are larger than normal. Thus the resonances are more damped.

A considerable amount of work has gone into theoretically predicting the positions of nasal resonances and antiresonances as well as spectrally analyzing nasal consonants from real speech. (Fujimura, 1962, Nakata, 1959). Fujimura (1962), making some simplifying approximations, derived positions of a single zero and four poles which represented the lowest resonances of the vocal tract. Figure 4.2 shows both the vocal tract configuration and an equivalent idealized representation for the consonant N. Fujimura calculated the transfer function (ratio of volume velocity at the extremity of the nasal cavity to volume velocity at the glottis) for N. This transfer function exhibited both poles and zeros. The quantities Bi, Bp, and Bm of Fig. 4.2 represent the susceptances (imaginary parts of the admittances) measured in the directions shown in Fig. 4.2. The zeros of the transfer function (assuming no loss) occur at frequencies where $Bm = \infty$, i.e., when the oral tract (acting as a side branch resonator here) shunts out the input. The poles of the transfer function occur at frequencies where $Bm + Bn + Bp = 0$, or $-Bm = Bn + Bp = Bi$. These are the natural frequencies of the configuration shown in Fig. 4.2. Figure 4.3 shows the pole and zero positions for articulatory configurations approximating those for the consonants N and M, as predicted by this technique. As seen in Fig. 4.3 the five lowest singularities of the transfer function are four poles and one zero for both M and N. These theoretical values for pole and zero positions provided

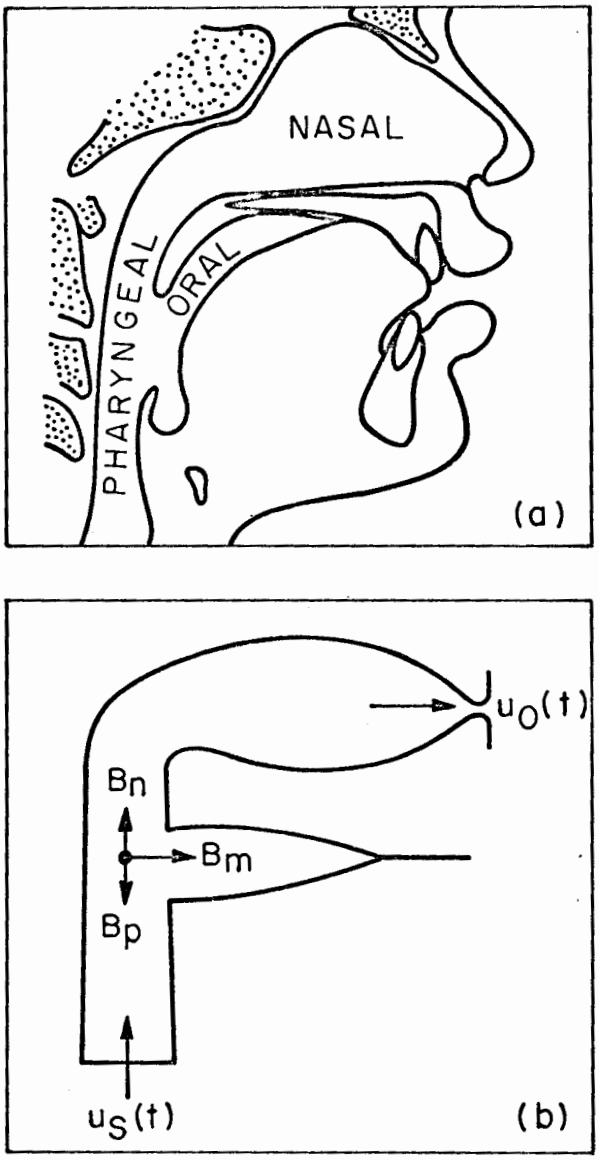


FIGURE 4.2

(a)-MIDSAGITTAL SECTION OF THE VOCAL AND NASAL TRACTS FOR /n/.

(b)-SCHEMATIZED ARTICULATORY STRUCTURE.

(FROM FUJIMURA - 1962)

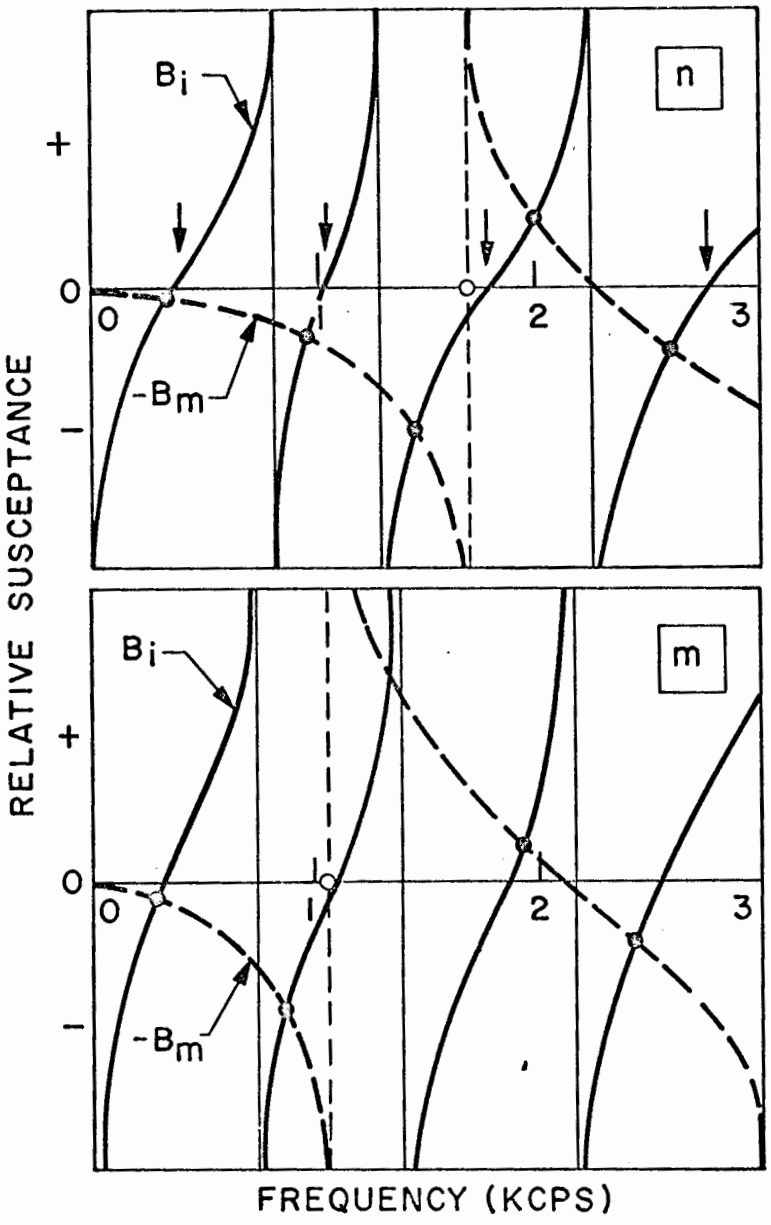


FIGURE 4.3
FREQUENCY CHARACTERISTICS OF SUSCEPTANCES B_i AND B_m
(SOLID CIRCLES INDICATE FORMANTS, OPEN CIRCLES INDICATE ANTI-FORMANTS.)
(FROM FUJIMURA - 1962)

excellent fits to observed spectra. Figure 4.4 shows a typical spectrum from real speech for the consonant M (Nakata, 1959).

From the above discussion we see that one way to synthesize the nasals is to cascade a variable pole (nasal pole) and a variable zero (nasal zero) with the three variable poles. At the same time we must keep in mind that the lowest pole has a broad bandwidth. This is the strategy we have adopted. The positions for the inserted nasal pole and nasal zero are:

	NASAL POLE		NASAL ZERO	
	Center Frequency (Hz)	Bandwidth (Hz)	Center Frequency (Hz)	Bandwidth (Hz)
M	1300	100	1100	100
N	1100	200	1700	200
NG	1000	200	2000	800

The three formant target positions are shown in Table 4.3. Table 4.3 presents the formant target positions for all consonants and the data in this table will be referred to in the following sections.

Nakata (1959) has shown, through a series of experiments with synthetic speech, that a nasal pole and zero need not be inserted to produce perceptable nasals. His strategy was to broaden the lowest resonance to 300 Hz bandwidth and this was sufficient to produce a nasal.

One side effect should be noted here. Our synthesizer has the nasal pole and zero in cascade with the other resonators. When a nonnasal sound is being produced the nasal pole and zero are set to identical values (1400 Hz) and their mutual effects are cancelled. When a nasal consonant is to be produced, this pole and zero must split. When moving the nasal pole from the rest position to its proper position for the nasal,

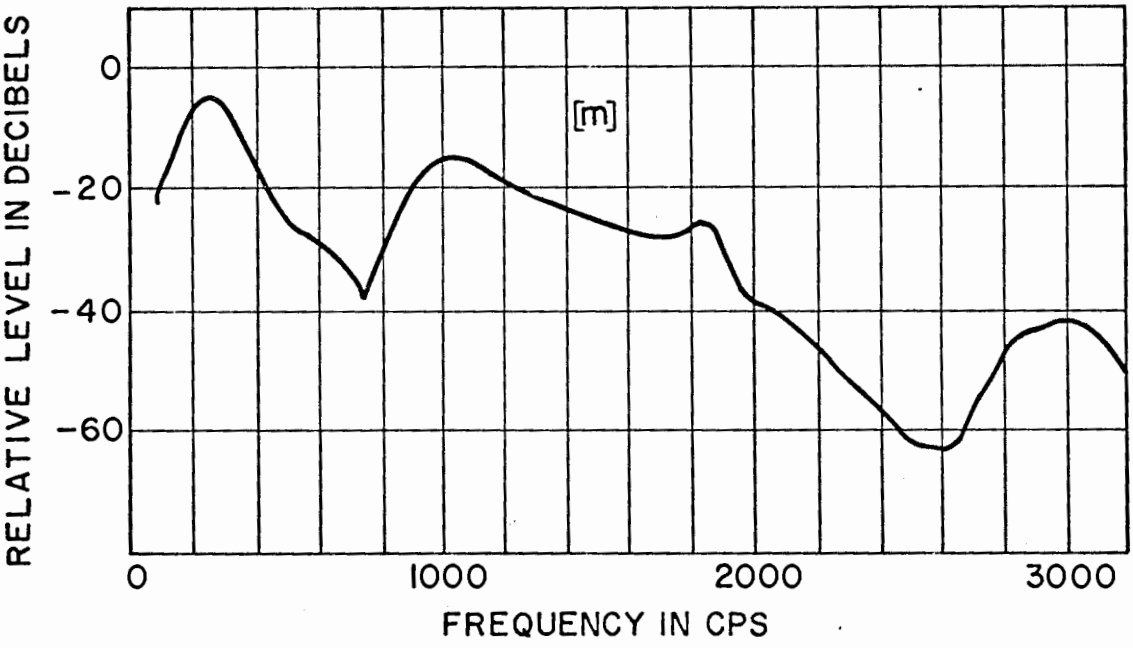


FIGURE 4.4
SPECTRAL ENVELOPE FOR A TYPICAL /m/.
(FROM NAKATA - 1959)

TABLE 4.3

CONSONANT FORMANT TARGETS

CONSONANT	F1 (Hz)	F2 (Hz)	F3 (Hz)	B1 (Hz)
B	0	800	1750	60
D	0	1700	2600	60
G	0	2350	2000	60
P	0	800	1750	60
T	0	1700	2600	60
K	0	2350	2000	60
M	280	900	2200	150
N	280	1700	2600	150
NG	280	2300	2750	150
F	175	900	2400	60
TH	200	1400	2200	60
S	200	1300	2500	60
SH	175	1800	2000	60
V	175	1100	2400	60
THE	200	1600	2200	60
Z	200	1300	2500	60
ZH	175	1800	2000	60

NOTES:

1. B2 = 100 Hz, B3 = 120 Hz for all consonants.
2. Stop consonant data from Delattre, 1958.
3. Nasal consonant data from Fujimura, 1961 - and Nakata, 1959.
4. Fricative consonant data determined from data of Lehiste and Peterson, 1960.

care must be taken to prevent the nasal pole from getting too close to the position of formant two or three. (For all practical purposes, within 200 Hz.) (In a cascade of resonators, if two resonance frequencies get too close, the response in the region of the resonance frequencies is very peaked, producing an undesirable audible squeal.) Unfortunately the nasal pole tries to cross the path of formant two quite often. There is no clear way to alleviate this problem. The solution which eliminates the squeal, invariably produces a somewhat less undesirable pop or burble. There is a good solution to this problem although it was never adopted. That is to provide an additional branch to the synthesizer for nasal sounds. In this way poles and zeros could be moved at will (since there would be no energy present in the nasal network until poles and zeros were at their proper positions) with no undesirable side effects. This technique has been used successfully by Fant in OVE II.

4.6 Fricatives

4.6.1 Voiceless Fricatives

The voiceless fricatives F, TH, S, and SH are not produced with the vocal cords vibrating. They are produced by a turbulence source applied at a point forward in the vocal tract. In the production of these sounds there is a high degree of constriction somewhat along the vocal tract. The turbulent source for voiceless fricatives is spatially located either at or somewhat forward of this constriction. Thus we have the case where we have a back cavity (from the open glottis to the constriction), a source of excitation, and then a front cavity. The transfer function will thus exhibit both poles and zeros.

Since the source for fricatives is not applied at a point but is instead spatially distributed, it is difficult to theoretically determine the placement of poles and zeros without further information as to source characteristics. However approximate models (Heinz and Stevens, 1961) have provided estimates of positions of the most important poles and zeros, and these have been verified closely by spectral analyses of speech.

For synthesis purposes the voiceless fricatives are spectrally divided into two groups. The first group contains S and SH. For both S and SH the spectrum can be adequately represented by a single pole and a single zero. Furthermore perceptual experiments (Nakata, 1960, Heinz and Stevens, 1961) have shown that samples synthesized from a single pole and zero are highly identifiable as S or SH depending solely on pole-zero positions.

The second group, containing F and TH cannot be represented adequately by a single pole and zero. The spectra of F and TH are characterized by a series of minor peaks and valleys for frequencies up to about 6 kHz. Their spectra can be considered flat up to these frequencies. Furthermore synthetic stimuli synthesized from a single pole-single zero approximation are not highly identifiable. The major perceptual cues for distinguishing F from TH are not contained in the fricative spectrum, but instead in the transitions from the adjacent sounds. In fact, it is mainly the second formant transitions that distinguish between F and TH. High second formant transitions lead to TH sounds whereas low formant two transitions lead to F sounds.

The synthesis strategy for F and TH is identical to S and SH - i.e., they are approximated by a transfer function with a single pole and

a single zero and are excited by the turbulent noise source. This is done to identify F and TH as fricative sounds, apart from any other class of sounds. The pole and zero positions for the voiceless fricatives are:

	FRICATIVE POLE		FRICATIVE ZERO	
	Center Frequency (Hz)	Bandwidth (Hz)	Center Frequency (Hz)	Bandwidth (Hz)
F	6500	970	3250	870
TH	6000	970	4200	870
S	4850	760	2750	1100
SH	2480	500	1250	900

4.6.2 Voiced Fricatives

The voiced fricatives are V, THE, Z, and ZH. These sounds are unusual in that they have two sources for their production. The vocal cords are vibrating and thus one source is at the glottis. However, the vocal tract is constricted at some point forward of the glottis producing a turbulence source in the neighborhood of the constriction. Thus the spectra of the voiced fricatives, strictly speaking, can be broken into two distinct components.

The voiced fricatives can also be separated into two classes for our synthesis strategy. In the first class belongs Z and ZH. Synthetic versions of Z and ZH distinctly require both excitation sources to be recognized reliably. The voiced spectrum is an all-pole function as one would expect. As discussed previously, the noisy part of the spectrum can be represented by a source which is the modulation product of a turbulence source and suitably shaped pitch pulses exciting a single pole-single zero transfer function. The pole and zero positions for Z and ZH are identical to those for S and SH.

V and THE, similar to F and TH, do not seem to rely heavily on the spectral properties of the noisy spectrum for their identifiability. In fact, it is quite often impossible to detect any unvoiced spectrum for V and THE on spectrograms. This is because the lower frequency voiced components dominate the spectrum. Thus our synthesis strategy for V and THE is to totally disregard all but the voiced spectrum. Thus V and THE are characterized entirely by positions of three formants.

The data of Table 4.3 shows formant positions for all the fricative consonants.

4.7 Stop Consonants

4.7.1 Voiced Stop Consonants

The voiced stop consonants are B, D, and G. They are produced by totally constricting the vocal tract at some point thus cutting off entirely the passage of air. Thus in terms of their static properties, the voiced stops are trivial to characterize. They can be synthesized by turning off everything. This is not strictly true for the voiced stop consonants. Often there is a small amount of low frequency energy radiated through the throat. This occurs when the vocal cords are able to vibrate (thus building up pressure tending to make them stop vibrating) even though the vocal tract is closed at some point. However, the point of interest concerning the stop consonants is their dynamic nature. The fact that a stop gap with silence or a voicing bar exists merely signifies that a stop consonant is being produced. It gives little or no information as to which stop consonant it is.

The perceptual attributes enabling man to distinguish between the various stop consonants are contained in the transitional region into the stop from the preceding sound, and the transitional region to the following

sound. An extensive series of experiments at Haskins Laboratories and at MIT has shown that transition timing and the motion of formants two and three serve to distinguish between B, D, and G. (Fujimura, 1961; Liberman et al., 1956; Delattre et al., 1955; Liberman et al., 1954.) Further, Fujimura (1961) has shown that a burst of noise immediately following the release of a G helps to distinguish a G from a D in many cases.

Formant frequencies used for synthesizing B, D, and G are shown in Table 4.3. Since there is no excitation during the stop gap, these frequencies represent virtual targets whose importance is realized primarily for transitional purposes. Thus a first formant frequency of 0 Hz represents the D.C. resonance of an ideal vocal tract closed at both ends.

Table 4.3 shows the virtual second formant resonance is low for B, and mid-valued for D and G. The third formant resonance for D is high, whereas for G it is mid-valued. It is these characteristics that provide a primary means of distinguishing between B, D, and G.

In producing a G or a K the point of articulation is dependent on whether the following vowel is a front or back vowel. We would expect the characterization of G and K to reflect this result. We will show in Chapter Five how this type of effect can be treated in the synthesis strategy.

4.7.2 Voiceless Stop Consonants

The voiceless stop consonants are P, T, and K. These consonants are produced in a manner similar to their voiced counterparts B, D, and G with one major exception. The duration of silence, or stop gap, is one of total silence. There is never a voicebar present. Following the

duration of closure (silence) there is a brief period of frication followed by a period of aspiration and then voicing begins.

Similar to their voiced counterparts, P, T and K are distinguished by transitional information and the information immediately following their release. During the silence of a voiceless stop consonant, pressure is built up in the mouth behind the closure. When the constriction is removed, this air behind the constriction is suddenly released causing a click like sound to occur. This sudden release of pressure is followed by a short period where the air released from the mouth is turbulent. This turbulent air (frication noise) is generated by the volume of air passing through the somewhat constricted vocal tract in the neighborhood of the original total constriction. This frication period is generally from 10 to 30 msec in duration. Of course the frequency spectrum of the frication noise is dependent highly on the place of the constricted, and hence the consonant. The frequency, duration, and amplitude of the frication noise for the voiceless stop consonants is seen below. (A_N will be explained in Chapter Five.)

		FRICATIVE		
Consonant	Frication Duration	Pole	Zero	A_N
P	5 msec	1450 Hz	725 Hz	30
T	10 msec	4200	2100	30
K	20 msec	2000	1000	15

(These data are from Bruce, 1966 and informal measurements made by the author.) The frequency positions of the pole and zero for K are variable (as is the point of articulation for K). When a vowel follows a K, the pole begins at a frequency just above the initial second formant frequency

for the vowel. This is in agreement with perceptual experiments (Liberman, Delattre, Cooper, 1952). Otherwise the pole position for K is as shown above.

The vocal cords do not begin vibrating, in general, following this frication phase. Instead the glottis is fairly wide open and air rushes through it. The articulators begin moving to positions appropriate for the following sound. Thus the second output during this period is that of aspiration. In terms of the synthesizer configuration we have a noise source exciting the formant network instead of the pulse source. Gradually (within 20-100 msec) the vocal cords begin to vibrate and we are in the transitional phase to the following sound. Perceptual experiments have shown (Guelke, Smith, 1963 and others) that the majority of the information for distinguishing between P, T, and K is contained in the fricational and aspirational period. There are other perceptual cues (such as formant transitions during the voiced portion of the transition) but these seem to be of secondary importance.

Table 4.3 shows the target formant values for P, T, and K. These are identical to those for B, D, and G as one might well expect.

4.8 Affricates

The affricates are CH and J. The production of CH is similar initially to T in that there is a stop gap followed by a burst of noise. However, immediately following the burst there is a long period of fricative noise with a spectrum similar to that of SH. There has been little study of the acoustic properties of CH, perhaps due to its low frequency of occurrence in American speech. (According to Denes, 1963, J occurs 0.51% of the time and CH occurs 0.37% of the time in spoken

English.) Therefore we have considered CH as a two phoneme sequence. The two phonemes are T and SH.

Similar statements may be made about J. Here we treat J as a tandem of D and ZH. An analysis of these phonemes must be made before a good acoustic approximation can be formulated.

4.9 H

The phoneme H is produced by air flowing through a fairly wide glottis and through the vocal tract. The characteristics of H are invariably those of the following vowel. That is, we can produce an H sound by exciting the synthesizer with noise instead of pitch pulses, while moving the formants to positions appropriate for the vowel.

4.10 Internal Open Juncture or Word Boundary

When people speak or read text out loud they appear to assimilate the information on a word by word basis. The acoustic waveform so attained is continuous in that there are not clear cut regions of silence between words. However it seems clear that word boundaries, or spaces between words have an effect on the acoustic waveform of an utterance.

People are able to distinguish accurately between phonetically equivalent pairs such as (grade A, gray day) or (an ice man, a nice man) when presented out of context. Therefore in pronouncing each of these pairs use must be made of the positions of the word boundaries. Lehiste (1959) made an acoustic-phonetic study of a group of such ambiguous pairs. She sought, by means of spectrographic analysis primarily, to determine the acoustic correlates of word boundary (called also internal open juncture) on phonemes near the juncture. Her primary results were the following. (Adopting Lehiste's convention, we shall call the phoneme

preceding the juncture a final allophone, and the phoneme following the juncture an initial allophone.)

1. Initial consonants are greatly lengthened in duration.
2. Voiceless stop consonants following final S are aspirated, whereas voiceless stop consonants following S in initial position or medial position are unaspriated. This distinction enables one to differentiate easily between such ambiguous pairs as (it sprays, it's praise), or (keep sticking, keeps ticking).
3. The formant positions for initial and final L differ significantly from each other and from medial L. (Similar results apply to R.) Furthermore, there are intensity and transition differences between initial and final L and R.
4. Initial vowels show gradual rises in intensity whereas final vowels show decreases in intensity.
5. Final vowels are lengthened.

Lehiste pointed out that the above results were generally applicable. There were other noticeable effects but their applicability seemed to be limited to specific cases.

We have tried to use these results in our synthesis scheme as they clearly play an important role in determining the patterns of speech.

CHAPTER FIVE

SYNTHESIS STRATEGY

5.1 Introduction

In Chapter Four we discussed the units in which the input information is coded - i.e. the phonemes. In our examination of the phonemes we talked of such entities as formant positions, pole and zero positions etc. These entities represented the quasi-static description of the phonemes. If speech were merely a succession of phonemes, with no interaction between successive phonemes, than this static description (with suitable modifications for bridging phonemes) would serve well. Unfortunately this is not the case. There are periods when the articulation is influenced by a single phoneme but generally the articulatory activity, and the associated acoustic waveform, is the result of more than one phoneme. It is the dynamic aspects of speech production that we are trying to model in the synthesis strategy.

In this chapter the synthesis strategy will be discussed in detail. This strategy has evolved over many months of experimentation. (Chapter Six presents a discussion of some of the experimentation.) The strategy described in this chapter represents the "current" status of our implementation. The data of both this chapter and Chapter Four is the data used in the current implementation. The examples of the synthetic speech that are referred to in Chapter Six and Seven are synthesized using the current implementation.

5.2 Information Rates

Before proceeding to an explication of the synthesis strategy, we shall take a brief look at the problem from an information theory point of view. A person speaking at an average rate will speak ten phonemes per

second. There are about thirty-two possible phonemes so (considering all phonemes equally probable) each phoneme contains five bits of information. The average rate of input information is thus fifty bits per second. As a measure of the output information we will use the accepted figure of fifty thousand bits per second for high quality speech (Flanagan, p4). The output information rate is about three orders of magnitude higher than the input rate. These figures clearly point out the burden which will be put on the scheme which converts from the phonemes to speech. It would be difficult for any scheme, from phonemic information alone, to produce speech that preserved anything but the intelligibility of the utterance. Features of speech, such as a speaker's accent, his rhythm, his speed, his emotional state, etc., could not be preserved without additional information.

Thus any scheme for synthesis by rule should be judged primarily on whether a native speaker of English recognizes and accepts an utterance as a possible utterance of the language. Intelligible speech is clearly our major goal. Of second most importance are his comments as to the "naturalness" of the utterance. Any further attributes of speech that are well reproduced, such as a perceived emotional state of the speaker, (in this case a computer), are strictly coincidental. We do not mean to imply that such effects cannot be included in our scheme. Additional input information (both linguistic and personal) would have to be included and rules for using this information would have to be devised. Whether such rules can be found or not is unknown at present. In any case we will not be concerned with such attributes of speech.

5.3 Brief Description of Synthesis Strategy

The input to the synthesis strategy is a linguistic description of the utterance. Included in this description are phonemes, word boundaries, vowel stress marks, pauses and a terminal marker indicating the end of the utterance.

For each phoneme which is a possible input a characterization is defined. This characterization contains the features of the quasi-static representation of Chapter Four, as well as certain other phoneme attributes. Amplitude and durational characteristics are also defined for the phonemes. Furthermore, for each phoneme a set of frequency regions around the formant target positions is specified. These frequency regions, in conjunction with the method of generating formant data, are used to define intrinsic phoneme durations.

The basic control of timing of events is derived from the formant data. Formant motion (i.e. the transitions of formants between phonemes) is described by the solution to a critically damped second degree differential equation. Transition time constants for each formant between each pair of phonemes are defined in matrices. The transition of any formant can begin before or after the transition of any other formant. The transition to a phoneme is terminated when all formants are within the frequency regions of that phoneme.

When the transition to one phoneme is terminated, the transition to the next phoneme can begin. In some cases the transition to the next phoneme does not begin until certain conditions are satisfied - e.g. a duration of steady state is generated for stressed vowels.

The motion of the remaining synthesizer control parameters is time-locked to the motion of the formants. The nasal and fricative poles and zeros initiate motion to their respective targets at the time that formants initiate transitions to new phonemes. The motion is linear and occurs within a time duration which is a function of the formant transition time constants. The source amplitudes are switched at times when the nasal and fricative poles and zeros are properly positioned.

The fundamental frequency contour is generated from input information and is also time-locked to the formant motion. A model for predicting fundamental frequency is postulated. The controlling variables are subglottal pressure and laryngeal tension. Archetypal contours for these variables are postulated. These contours are modified by stressed vowels, voiced and voiceless consonants, the sentence boundary marker and an indication of whether or not the sentence is a yes-no question. The times during which these modifications take place are determined from the formant data.

5.4 General Framework and Details of Synthesis Strategy

There are two levels on which we must attack the problem of synthesis by rule. The higher level embodies a framework of the essential ideas for producing speech. The higher level parallels in many ways the ideas of our predecessors who worked on speech synthesis by rule. It will be more general than any previous scheme and will contain many original ideas.

Along with the higher level ideas will be the lower level implementation which we have adopted. Specific techniques and explicit data are presented. These techniques and data, along with the data of Chapter Four, serve to define the current synthesis strategy.

Throughout this chapter we must keep in mind the three basic constraints on our scheme. These are:

1. A phonemic input
2. An acoustic domain approach
3. A terminal analog synthesizer as the output device.

5.4.1 Phoneme Characterization

Each phoneme has a unique characterization independent of adjacent phonemes. This characterization, for the most part, embodies all the static properties of the phonemes that were discussed in Chapter Four. It tells the synthesizer how to set parameters to produce the best possible version of the phoneme. In our scheme this characterization will include acoustic parameters, or quantities that are easily converted to acoustic parameters. However for any synthesis from phonemes a characterization would be necessary. For instance an articulatory scheme would require, for each phoneme, data such as positions of the constrained parts of the tongue, mandible position, lip position etc. Thus the first constraint above, a phonemic input, requires that we have phonemic characterizations.

The characterization necessary for our acoustic domain synthesis includes formant information, source characteristics in the production of the phoneme, a description of whether it is nasal or fricative, and a set of frequency regions surrounding the formant positions.

The formant information is a set of target positions for both center frequency and bandwidth of formants one, two and three. These target positions are necessary for situations when a steady state is reached, as might occur in stressed vowels. For a consonant they sometimes (as in the case of F or S) represent "virtual" steady state

frequencies even though these resonances may never be excited during the realization of the phoneme. The formant target positions correspond to the lowest three resonances of the vocal tract configuration appropriate to the phoneme.

The source characteristics describe the condition of the vocal cords during the production of the phoneme. (Hence they describe whether or not a turbulent source is used to produce the phoneme.) If the vocal cords are vibrating the sound is voiced. (Vowels, nasals, voiced fricatives, and voiced stops are voiced.) Unvoiced sounds are characterized by a vocal tract configuration which is highly constricted in some region. In the neighborhood of this constriction, the air stream coming from the glottis becomes turbulent. Thus the source for voiced sounds is at the glottis, whereas for unvoiced sounds it is in the neighborhood of the point of maximal constriction of the vocal tract. Voiced fricatives, although technically considered voiced sounds, are both voiced and unvoiced in that there is a point of constriction in the vocal tract acting as a further source of excitation.

Nasality characteristics describe the position of the velum, in an abstract way, during the phoneme production. Fricative characteristics must be given to distinguish between the two classes of voiceless sounds - voiceless stops, and voiceless fricatives.

The frequency regions of a phoneme represent a "semidynamical" part of the characterization. Loosely speaking, they represent the degree to which certain acoustic parameters must approximate the target values of these parameters in order for the phoneme to be realized. In an articulatory analog the corresponding concept would be to what extent must a

given vocal tract configuration approximate the target configuration for the phoneme. For instance a major articulatory feature of the stop consonant B is total closure at the lips. The position of the back of the tongue is unconstrained so far as B is concerned. Thus a configuration with closure at the lips and any position of the back of the tongue could be considered sufficiently close to the target configuration for B. A better example would be the case of the high front vowel IY. During the production of an IY, the tongue assumes a certain position with the front part high in the mouth. A configuration with the front of the tongue in the vicinity of the position for an IY would be considered acceptable for an IY - especially in the context of connected speech.

Thus the frequency regions represent a compromise between choosing a single characterization for a phoneme (either acoustic or articulatory) and considering it inviolate, and the realization that there are many acceptable characterizations for a phoneme - especially in the context of connected speech.

In the acoustic domain the frequency regions represent regions around the target formant positions through which the formants must pass for the phoneme to be realized. Thus the degree of approximation is in terms of the lowest three formants (resonances), corresponding to approximation in the articulatory domain in terms of vocal tract configurations. (Two vocal tract configurations differing by a small amount, will yield two sets of resonance frequencies differing similarly by a small amount.)

The extent of the frequency regions is a measure of the minimum necessary degree of approximation. (It will be shown later in this chapter that the smaller the frequency regions for a phoneme, the longer its

intrinsic duration.) For instance a tense vowel might be given smaller frequency regions than its lax counterpart indicating that the formants for a tense vowel must be close to the target values to realize such a vowel whereas a lax vowel can be perceived correctly even though the formants are not as close to the target values.

The frequency regions also play a large role in determining the timing of events in connected speech. This aspect of the frequency regions will be discussed in a later section.

A list of the phoneme characterizations we have used is shown in Table 5.1. The first three columns in Table 5.1 list the formant target positions of the phonemes. We have already discussed these data and their sources in Chapter Four. The second three columns show the frequency regions of the phonemes. The frequency regions for the phonemes were chosen from both experimental and theoretical results. For vowels, the frequency regions were chosen so the characteristic duration of an unstressed vowel would be within the correct range of values. For W, L, R, and Y, similar considerations were used. For nasals, the frequency regions were used to assure a minimum nasal duration. The frequency regions for the fricative consonants were made small to produce long intrinsic fricative duration. This was done because it was found, during the course of this thesis work, that long fricatives generally increased the overall intelligibility of an utterance. The final three columns of Table 5.1 describe nasality, fricative and voicing characteristics of the phonemes. A + in any column indicates the presence of the feature and a - indicates the absense of this feature. The voicing condition of the voiced fricatives Z and ZH is ± indicating the two sources used to produce these sounds on the synthesizer.

TABLE 5.1

PHONEME CHARACTERIZATIONS

PHONEME	F1	F2	F3	Δ1	Δ2	Δ3	NASAL	FRIC.	VOICED
IY	270	2290	3010	75	75	150	-	-	+
I	390	1990	2550	75	75	110	-	-	+
E	530	1840	2480	75	80	110	-	-	+
AE	660	1720	2410	75	75	110	-	-	+
UH	520	1190	2390	75	75	75	-	-	+
A	730	1090	2440	37	75	115	-	-	+
OW	570	840	2410	75	75	115	-	-	+
U	440	1020	2240	75	75	90	-	-	+
OO	300	870	2240	75	80	90	-	-	+
ER	490	1350	1690	75	80	100	-	-	+
W	300	610	2200	25	40	150	-	-	+
L	380	880	2575	25	80	150	-	-	+
R	420	1300	1600	30	80	100	-	-	+
Y	300	2200	3065	25	110	200	-	-	+
B	0	800	1750	50	75	120	-	-	+
D	0	1700	2600	30	50	160	-	-	+
G	0	2350	2000	15	50	100	-	-	+
M	280	900	2200	17	17	40	+	-	+
N	280	1700	2600	17	17	100	+	-	+
NG	280	2300	2750	17	17	100	+	-	+
P	0	800	1750	50	40	80	-	-	-
T	0	1700	2600	30	30	100	-	-	-
K	0	2350	2000	10	30	70	-	-	-
F	175	900	2400	20	34	80	-	+	-
TH	200	1400	2200	20	28	68	-	+	-
S	200	1300	2500	20	28	50	-	+	-
SH	175	1800	2000	10	34	100	-	+	-
V	175	1100	2400	10	15	100	-	+	+
THE	200	1600	2200	10	15	100	-	+	+
Z	200	1300	2500	20	30	50	-	+	±
ZH	175	1800	2000	10	40	100	-	+	±

(The data for Δ1, Δ2, Δ3 were determined experimentally.)

5.4.2 Durational and Amplitude Characteristics

Durational and amplitude characteristics of the phonemes can be considered dynamical in nature. Durational characteristics are dynamical in the sense that phoneme duration is a function of context and content of the utterance. Stressed vowels are longer than unstressed vowels whereas a stressed vowel preceding a B is longer than the same stressed vowel preceding a P. Durational and amplitude characteristics must be specified since the rate of change of amplitudes is of perceptual importance in identifying phonemes.

The essential durational features concern both vowels and consonants. Vowel duration must be specified for stressed vowels. (Unstressed vowels need not have durations specified as the dynamical aspects of the synthesis strategy will generate the correct durations for unstressed vowels.) The duration of a stressed vowel is a function of the following phoneme. The longest vowels are those followed by voiced fricative consonants, whereas the shortest are followed by voiceless stop consonants (House, 1961). Some algorithm or table must be specified to produce correct durations of stressed vowels.

W, L, Y, and R are similar to unstressed vowels. No durational characteristics need to be supplied for them as their duration is generally short and the mechanics of the speech generating scheme will generate acceptable durations.

For certain consonants maximum durations are specified. The dynamical aspects of the synthesis strategy will generate acceptable consonant durations in most cases. Consonant duration (as measured from human speech) is not a fixed quantity but is highly context dependent.

(For example initial consonants are much longer than medial consonants.) The synthesis strategy should generate consonants whose duration is variable within certain limits. Maximum durations need to be specified to prevent the consonant from being unnaturally long and hence objectionable.

For stop consonants the maximum duration of the stop gap is specified. (An unnaturally long stop gap will be heard as a pause or a break in the speech.) The maximum duration of the stop gap is variable with the place of articulation of the stop. (Velar stops have the longest stop gaps whereas labial stops have the shortest stop gaps.) For unvoiced stop consonants the duration of the frication noise is specified. (The noise duration is sufficiently short (5-15 msec) such that its variation can be neglected.) Furthermore, the duration of aspiration for voiceless stops, as a function of context, must be specified. Aspiration duration is longest when a stop is followed by W, Y, L, or R, and shortest when followed by an unstressed vowel.

For nasal consonants the maximum duration is specified. If the nasal duration is too short, it may be perceived as a voiced stop consonant instead. The synthesis strategy is structured to guarantee the nasal duration will exceed some desired value.

For fricatives no durational characteristics are included. When two fricatives follow each other (as in Father's shoe) the duration of the first is very short. Hence no minimum duration is specified. Furthermore, fricatives sounds are often greatly exaggerated in length (for emphasis) in connected speech. A speaker can and often will articulate a fricative sound for a long time as in the emphasis on s in "This book". Thus no maximum duration for fricative sounds is specified.

For the consonant H the duration of the aspiration is specified. The aspiration duration should vary with context. An H followed by a stressed vowel is longer than an H followed by an unstressed vowel.

Amplitude characteristics of the phonemes are specified. For sounds which are voiced, the level of voicing amplitude (A_V in our synthesizer) is specified, whereas for voiceless sounds, noise amplitude (A_N) must be specified. Furthermore, the rate at which the amplitudes of noise and voicing are varied is of great importance. For stop consonants the amplitudes of noise and voicing are set to zero essentially instantaneously, indicating a rapid total closure. Source amplitudes change to their correct values almost instantaneously following the release for stops.

Amplitude characteristics for vowels differ greatly from stops. There is always a gradual buildup of voicing amplitude and similarly a gradual decay following the vowel. Similar results hold for W, L, R, and Y, and the nasal sounds.

For fricative sounds the buildup of noise level is of importance. There is a more rapid buildup of noise level for voiceless fricatives than for their voiced counterparts. This effect is included in our synthesis strategy.

A modification of House's data on durations of stressed vowels was used in our strategy. (Table 4.2 shows House's original data.) Informal experiments by the author have shown that the durations of stressed vowels are unnaturally long using House's data. This is due to the fact that the environment in which the stressed vowels occurred in House's experiment does not parallel the environment of stressed vowels

in real speech. The data of Table 4.2 was modified for use in the synthesis strategy by subtracting 100 msec from all values in the table.

For stressed diphthongs, the durational characteristics are manifested in one of two ways. For /e^{I}/ and /o^{U}/, which are treated as combinations of E + IY and OW + OO, increases in duration are reflected in increased duration of the second phoneme of the sequence. The first phoneme of the sequence never has its duration specified and hence its duration is independent of whether the diphthong is stressed or not.

For the diphthongs /aI/, /ɔI/, and /aU/ the case is somewhat different. Whenever these diphthongs are stressed, the duration of neither phoneme of the sequence is increased. Instead the duration of the transition between the first and second phonemes is increased. This effect has been observed by the author on spectrograms of real speech. In a sense, the diphthongs /aI/, /ɔI/, and /aU/ are better characterized by their transitional features than by steady state resonance at a pair of target configurations.

For the stop consonants the maximum durations of the stop gap are:

60 msec. for B and P

80 msec. for D and T and

100 msec. for G and K.

The durations and amplitudes of the frication noise for the stop consonants were specified in Chapter Four. The aspiration durations for voiceless stop consonants followed by unstressed vowels are:

40 msec. for P

60 msec. for T and

80 msec. for K.

(The amplitude control of the synthesizer (A_V) is set to 70 for aspiration.) These figures are increased by 25% for stop consonants followed by stressed vowels and by 50% when followed by R, L, W, or Y.

The maximum nasal duration is set to 150 msec for all nasals. This value is about the longest nasal duration observed in real speech.

Finally, the aspiration duration for an H is 110 msec and is independent of context. We suggested previously that the aspiration duration should depend on the nature of the following consonant but we have not used this approach in this implementation.

The amplitude characteristics of the phonemes are presented in Table 5.2. The columns headed A_N, A_V, and A_{VB} represent the levels of the amplitude controls during the phoneme production. (Hence a voiceless stop such as P has all amplitudes zero indicating a stop gap without a voice bar.) The columns headed DAVMS and DANMS indicate rates of change of the controls A_V and A_N during the time in which the source characteristics are changing. Hence for B the values of DAVMS and DANMS are 20 and 2. If the values of A_V and A_N are 100 and 30 prior to switching, it will take 5 msec for A_V to change to the value of 0 for a B, and 15 msec for A_N to reach 0. Hence it would take 15 msec for all sources to be turned off in this hypothetical case. For the stop consonants, the values of DAVMS and DANMS apply both to switching from the previous phoneme and switching to the following phoneme. Hence stop consonants will exhibit both rapid closure and rapid release in their production.

5.4.3 Formant Motion

One of the most important features of this implementation is the method of generating formant data. Previous schemes have used linear

TABLE 5.2

SOURCE AMPLITUDE CHARACTERISTICS

PHONEMES	A_V	A_N	A_{VB}	DAVMS	DANMS
B,D,G	0	0	1	20	2
P,T,K	0	0	0	50	2
M	65	0	0	1.5	2
N	55	0	0	1.5	2
NG	50	0	0	1.5	2
F	0	15	0	4	2
TH	0	28	0	4	2
S	0	40	0	4	4
SH	0	30	0	2	1
V	80	0	0	2	1
THE	50	0	0	2	1
Z	65	26	0	2	1
ZH	100	20	0	2	1
Vowels and W,L,R,Y	100	0	0	4	2

(These data were determined experimentally.)

formant motion (Haskins, Mattingly et al.) or combinations of linear and parabolic motion (Kelly and Gerstman). There is strong evidence that linear motion of formants is too restrictive. The synthesis of many consonant-vowel pairs, such as the voiced stop D followed by the vowel IY, cannot be adequately done with a linear transition. The transition must be less abrupt. (Informal synthesis experiments by the author and others have shown that multisegment piecewise linear approximations to such consonant-vowel pairs can produce a good syllable whereas a single segment linear approximation proved unsuccessful.) Smooth, continuous transitions are needed in such cases. Furthermore, formant transitions seen on spectrograms of speech are smooth and continuous.

In an effort to match these observed characteristics of formant transitions, the motion of formants in this implementation was described by the solution to a critically damped second degree differential equation. A second degree differential equation was chosen because it provided a good fit to data on formant transitions. The reason a critically damped solution was employed was that only a single time constant (the inverse of the double order real axis pole value) was necessary to characterize completely the response to a forcing function. Thus we could determine values for the time constants from examining formant transitions for real speech on spectrograms.

The differential equation used in this implementation is seen in Eq. (5.1).

$$\frac{d^2x(t)}{dt^2} + \frac{2}{\tau}\frac{dx(t)}{dt} + \frac{1}{\tau^2}\,x(t) = F(t) \qquad (5.1)$$

where

$x(t)$ = formant value as a function of time

τ = time constant of response

F = force function applied at input.

An electrical circuit which realizes Eq. (5.1) is seen in Fig. 5.1. The transfer function appropriate to Eq. (5.1) is expressed in Eq. (5.2).

$$\frac{x(s)}{F(s)} = H(s) = \frac{\left(\frac{1}{RC}\right)^2}{\left(s + \frac{1}{RC}\right)^2} = \frac{\frac{1}{\tau^2}}{\left(s + \frac{1}{\tau}\right)^2} \tag{5.2}$$

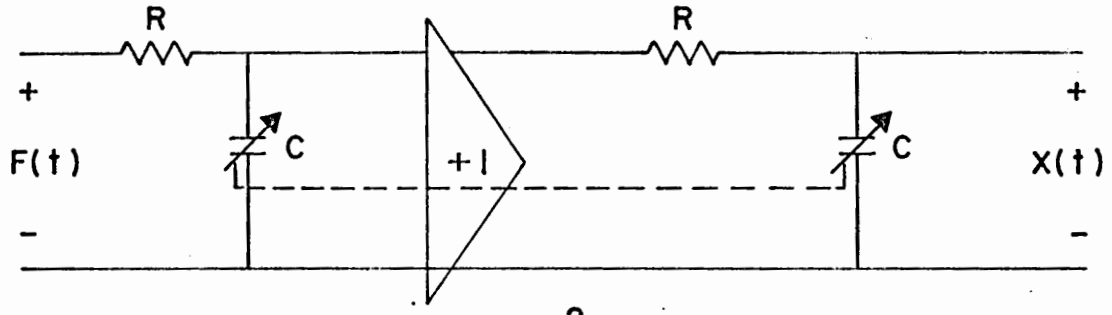
The impulse response of the system of Fig. 5.1 is seen in Eq. (5.3) (assuming initial rest conditions).

$$x(t) = \frac{t}{\tau^2} \exp(-t/\tau) u_{-1}(t) \tag{5.3}$$

The step response (again assuming initial rest conditions) is seen in Eq. (5.4).

$$x(t) = [1 - (1+t/\tau)\exp(-t/\tau)]u_{-1}(t) \tag{5.4}$$

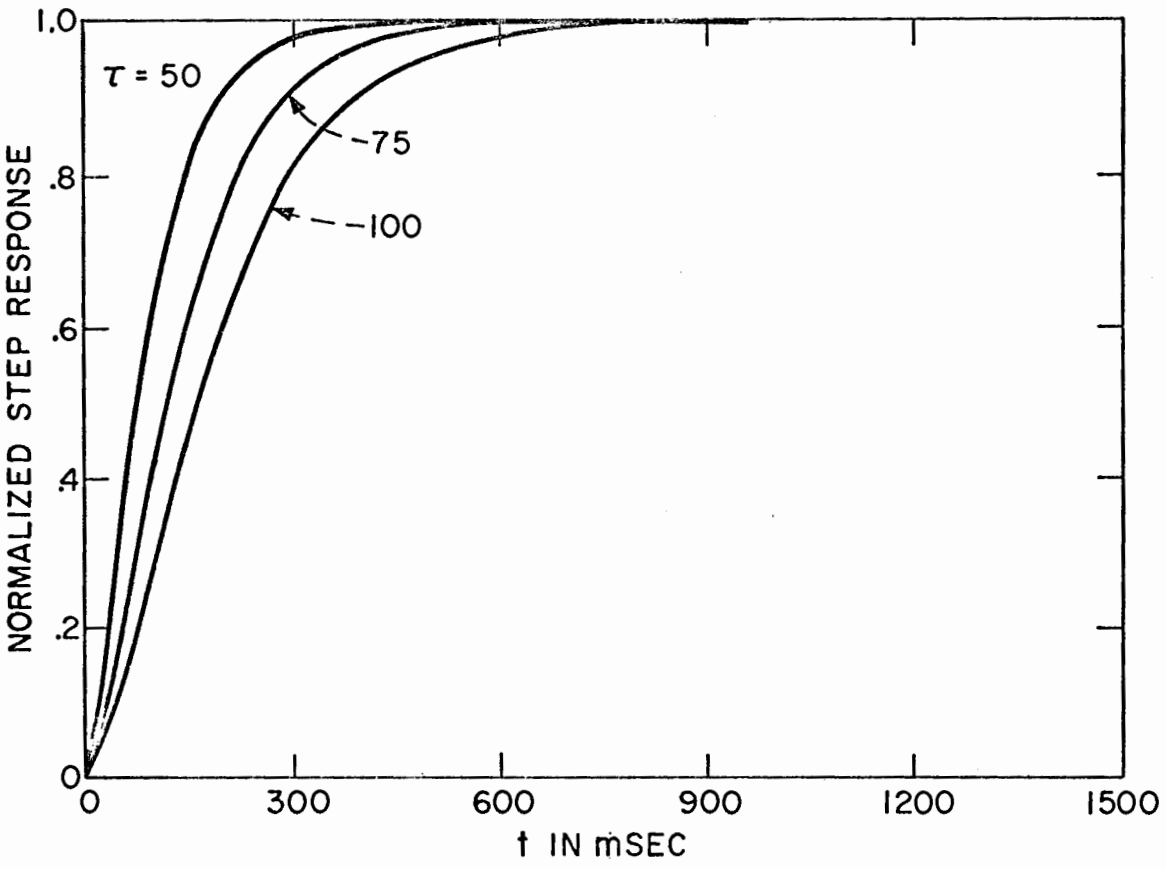
Figure 5.2 shows examples of the step response (suitably normalized) for different time constants τ.



$$\frac{X(S)}{F(S)} = \frac{\left(\frac{1}{RC}\right)^2}{\left(S + \frac{1}{RC}\right)^2}, \quad \tau = RC$$

FIGURE 5.1

ELECTRICAL REALIZATION OF EQUATION 5.1



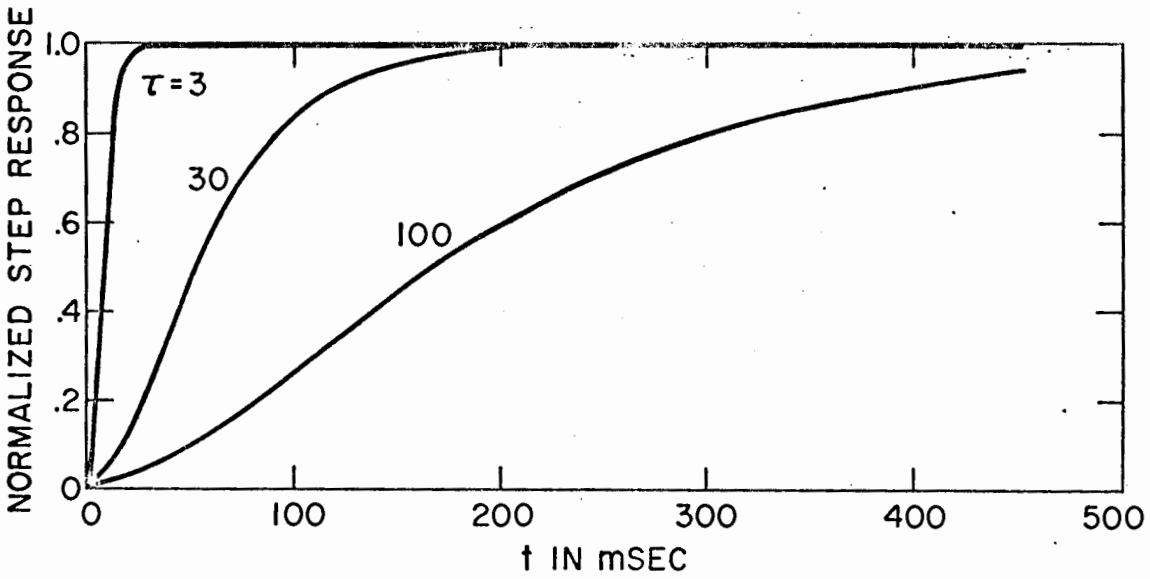


FIGURE 5.2

NORMALIZED STEP RESPONSES FOR CIRCUIT OF FIG. 5.1

To see how we would apply this technique for generating data on formant transitions we shall consider two cases. In case one we consider motion from a steady state at target 1 (formant value Ai) at time t = 0 to a steady state at target 2 (formant value Af) with a time constant τ. For $t < 0$ the input forcing function F(t) is at value Ai/τ^2 and thus the steady state formant value (x(t)) is Ai as postulated (as seen from Eq. (5.1)). At t = 0, F(t) changes in a step-like fashion to the value Af/τ^2 thus generating the overall formant response x(t) seen in Eq. (5.5).

$$x(t) = Af + (Ai-Af)(1+t/\tau)\exp(-t/\tau) \qquad (5.5)$$

As time t gets much larger than τ (say $t = 5\tau$) x(t) will be approximately equal to Af as desired.

In general, motion between target positions does not proceed from a steady state condition - i.e. there are initial conditions. Therefore in case two we consider motion to a target whose formant value is Af, from an initial formant position Ai with an initial formant velocity Vi $\left(\left.\frac{dx}{dt}\right|_{0^-}\right)$. The solution of Eq. (5.1) for such initial conditions is seen in Eq. (5.6).

$$x(t) = Af + (Ai-Af)\exp(-t/\tau) + \left[Vi + \frac{(Ai-Af)}{\tau}\right]t\ \exp(-t/\tau) \qquad (5.6)$$

$$\text{for } t \geq 0$$

At t = 0, the point in time when formant motion to the new target begins, both the formant value, and the formant velocity are continuous.

(Equation (5.6) trivially reduces to Eq. (5.5) for the case where Vi = 0, or no initial formant velocity.)

For a computer simulation of this method, Eq. (5.2) was z-transformed using impulse invariant techniques (to preserve the time response to an impulse input) giving Eq. (5.7) for the digital filter.

$$\frac{x(z)}{F(z)} = H(z) = \frac{(1-k)^2 z^{-1}}{(1-k\ z^{-1})^2} \tag{5.7}$$

where

$k = e^{-T/\tau}$

and

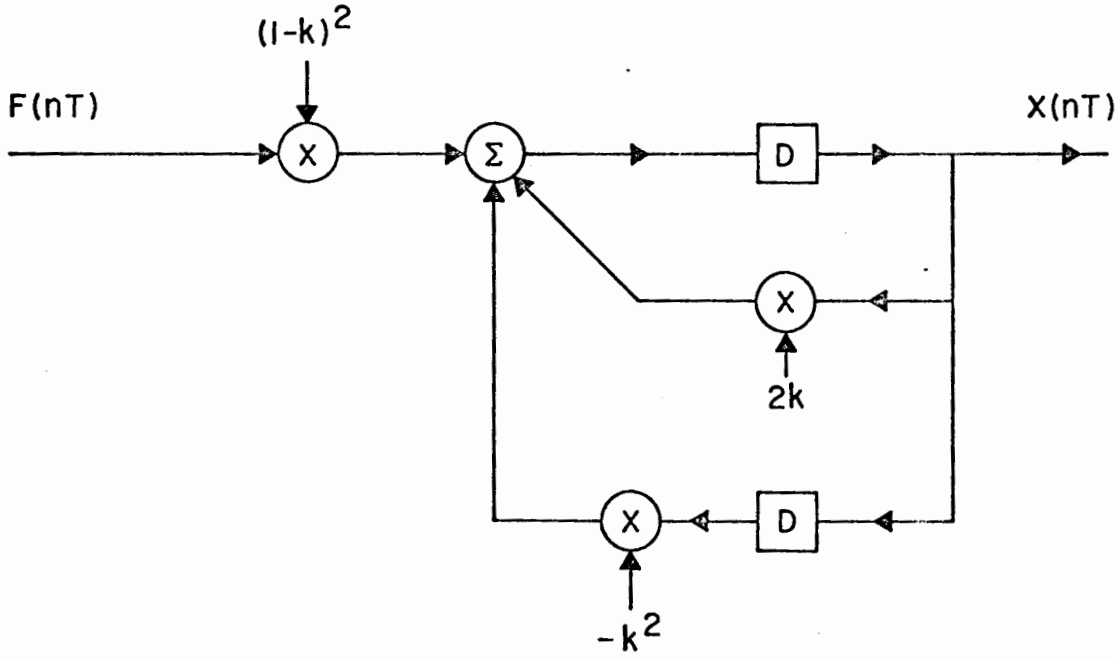
T = sampling time = 1 msec

τ = time constant in msec

(A sampling time of 1 msec was chosen so rapid formant transitions could be approximated using the synthesis strategy.) The difference equation used to generate the formant data is shown in Eq. (5.8).

$$x(nT) = 2kx(nT-T) - k^2 n(nT-2T) + (1-k)^2 F(nT-T) \tag{5.8}$$

The computer realization of this difference equation is shown in Fig. 5.3.



$$X(nT) = 2k\ X(nT-T) - k^2\ X(nT-2T) + (1-k)^2\ F(nT-T)$$

FIGURE 5.3

DIGITAL FILTER REALIZATION OF FORMANT DATA GENERATION NETWORK

The digital response of the network of Fig. 5.3 to a simulated step has been theoretically determined and is shown in Eq. (5.9).

$$x(nT) = 1+kne^{-nT/\tau}-e^{-nT/\tau}(1+n) \tag{5.9}$$

and the response to a simulated impulse is seen in Eq. (5.10).

$$x(nT) = ne^{-nT/\tau}(1-k)^2e^{T/\tau} \tag{5.10}$$

Since impulse invariance techniques were used to determine the filter, the impulse responses of the continuous and digital filters are identical at the sampling instants. Furthermore, the step responses of both the continuous and digital filters are within 1% of each other for all values of τ of interest. It is clearly important that the digital step response closely approximate the continuous step response as we will always be applying a step input to our filter.

We have dwelled on our technique for generating formant data as its importance to the rest of this implementation cannot be underestimated. It is the basic mechanism from which timing is controlled. It serves to define intrinsic phoneme durations. Source switching times and times at which nasal and fricative poles and zeros move will be defined in terms of the formant data. Hence a large part of our work has evolved around this technique.

5.4.4 Matrices of Time Constants

One of the most important aspects of an acoustic domain synthesis by rule is the transitions between phoneme target values. The transitional data is of extreme importance for perceiving consonants and the more closely we can approximate the formant transitions of real speech, the better the chances the consonant will be perceived correctly.

In the synthesis strategy formant transitions between two phonemes are a function of only those two phonemes. There is evidence (Ohman, 1966) that for certain vowel-consonant-vowel transitions there are coarticulational effects and hence three phonemes influence the formant transitions. We have not included these coarticulation effects to date.

Formants move at different rates during their transitions. This is a result observed from spectrograms of real speech. No thorough experiment has been conducted to test the perceptual importance of varying rates of formant transitions.

The data on formant transitions are presented in the form of matrices. These matrices contain transition time constants for motion from one phoneme to another. These time constants are those referred to in the previous section-- i.e. the ones which are used in Eq. (5.1). The source of the time constant data is spectrograms of real speech.

In the general case three matrices of time constants (one for each formant) are necessary. In the present implementation there are 31 phonemes as possible inputs. Thus each matrix involves 961 pieces of data or 2883 time constants overall. This is an exceedingly large amount of data. However this much data need not be specified. Certain approximations and phoneme groupings have reduced the number of time constants considerably.

The original group of 31 phonemes was subdivided into two groups. The first group contained the ten vowels - IY, I, E, AE, UH, A, OW, OO, U, and ER. The second group contained eleven subgroups of consonants which were partially arranged by place of articulation. These are: (B,P,M), (D,T,N), (G,K,NG), (F,V), (TH, THE), (S,Z), (SH, ZH), (W), (R), (L), and (Y). To a first approximation phonemes that are articulated at the same place and in the same manner should have identical time constants of motion to any other phoneme. Thus motion between formant targets for B to R should take approximately the same time as motion between formant targets for P to R. Spectrographic evidence is that the approximation is a good one.

At this point consider a single one of the three matrices. For purposes of example we shall consider the matrix for formant one. Initially we had a 31X31 matrix. It is now broken down in 4 matrices - a 10X10 matrix for vowel-vowel (V-V) transitions, an 11X11 matrix for consonant-consonant (C-C) transitions, an 11X10 matrix for consonant-vowel (C-V) transitions, and a 10X11 matrix for vowel-consonant (V-C) transitions. Thus the total number of entires is 441 as opposed to 961 originally. To reduce this number further, certain simplifying assumptions were made. The assumption was made that the V-V and C-C matrices were symmetric - i.e. $\tau_{V1V2} = \tau_{V2V2}$; $\tau_{C1C2} = \tau_{C2C1}$. A further approximation was that the V-C matrix was derivable from the C-V matrix. For V-C pairs where C was a stop or nasal, the approximation used was that $\tau_{CV} = \tau_{VC}$, i.e. the time constant from the vowel to the consonant equalled the time constant from the consonant to the vowel. For all other V-C pairs, the approximation $\tau_{VC} = 2\tau_{CV}$ was used. These approximations are clearly ad hoc ones chosen to fit a large

percentage of the available data. No further justification for the fit can be given.

The number of entries in the matrix was further reduced since many of the combinations of two phonemes could never appear as a possible input - i.e. no utterance could contain the phoneme A followed by ER etc. Thus the total number of entries that had to be specified (measured from spectrograms) was about 150 as opposed to the original figure of 961. This is a more manageable figure.

The phoneme H, although an acceptable input symbol, is not included in our matrix. Transitions to H assume the time constants of transitions to the vowel which follows the H.

Further reductions in the total number of entries for the three matrices were made. For transitions between two vowels or two consonants, the time constants for all formants were made identical. Furthermore, the time constant for formant three was made equal to the time constant for formant two for C-V and V-C pairs for all consonants which were not stops or nasals. Thus a total of 300-350 numbers were sufficient to describe the transition matrices.

The matrices used in this implementation are shown in Tables 5.3a, b, c, d, e, f, g, and h. (Tables for V-C transitions are included to illustrate the cases where exceptions to the approximations were made.)

5.4.5 Delay of Initiation of Formant Transitions

The point in time at which one formant initiates motion to a new target value need not coincide with the time at which another formant similarly begins to move. To describe this effect, in a general case, there must be a set of delay matrices describing delay of motion of each

TABLE 5.3a – CONSONANT-CONSONANT TRANSITION MATRIX

	B P M	D T N	G K NG	F V	TH THE	S Z	SH ZH	W	L	R	Y
B,P,M	5	10	10	10	20	30	10	10	22	7	10
D,T,N	10	5	10	10	10	30	10	35	10	25	10
G,K,NG	10	10	5	10	6	25	10	10	25	23	10
F,V	10	10	10	10	10	25	10	10	35	3	10
TH,THE	20	10	6	10	10	10	10	10	10	15	10
S,Z	30	30	25	50	10	10	10	22	30	10	10
SH,ZH	10	25	10	10	10	10	10	10	10	30	10
W	10	35	10	10	10	22	10	10	10	10	10
L	15	3	25	35	10	40	10	10	10	10	10
R	7	25	23	3	15	10	30	10	10	10	10
Y	10	10	10	10	10	10	10	10	10	10	10

THE CONSTANTS FOR F_1, F_2, F_3 IN MILLISECONDS

TABLE 5.3b — VOWEL – VOWEL TRANSITION MATRIX

	IY	I	E	AE	UH	A	OW	U	OO	ER
IY	20	20	37	20	20	25	23	20	20	20
I	20	20	20	20	20	20	20	20	20	20
E	37	20	20	20	20	20	20	20	20	20
AE	20	20	20	20	20	20	20	20	20	20
UH	20	20	20	20	20	20	20	20	20	20
A	25	20	20	20	20	20	20	20	18	20
OW	23	20	20	20	20	20	20	20	25	20
U	20	20	20	20	20	20	20	20	20	20
OO	20	20	20	20	20	18	25	20	20	20
ER	20	20	20	20	20	20	20	20	20	20

TIME CONSTANTS FOR F_1, F_2, F_3 IN MILLISECONDS

TABLE 5.3c – CONSONANT – VOWEL TRANSITION MATRIX

	IY	I	E	AE	UH	A	OW	U	OO	ER
B, P, M	30	8	16	10	13	15	10	16	13	13
D, T, N	16	19	33	33	33	33	16	16	10	33
G, K, NG	26	40	25	42	24	67	20	13	13	30
F, V	10	10	10	15	15	15	8	5	10	10
TH, THE	10	16	10	20	12	25	24	10	10	23
S, Z	6	6	16	20	17	27	13	13	6	20
SH, ZH	6	10	30	30	33	35	33	15	10	33
W	15	20	25	37	6	22	10	6	6	16
L	3	3	3	3	3	3	3	3	3	3
R	16	19	20	20	12	15	15	6	12	10
Y	15	12	16	16	24	13	20	18	16	24

TIME CONSTANTS FOR F_1 IN MILLISECONDS

TABLE 5.3d - CONSONANT - VOWEL TRANSITION MATRIX

	IY	I	E	AE	UH	A	OW	U	OO	ER
B,P,M	15	16	15	15	15	15	10	16	33	10
D,T,N	67	21	10	16	33	33	33	33	30	40
G,K,NG	26	42	28	40	35	33	30	16	35	30
F,V	26	20	13	45	19	10	5	5	25	10
TH,THE	33	33	30	30	12	10	16	22	20	23
S,Z	22	26	26	30	17	17	13	13	32	30
SH,ZH	25	25	15	15	40	35	33	33	52	33
W	15	20	25	37	6	22	10	6	6	16
L	18	12	18	21	6	9	6	6	9	15
R	16	19	20	20	12	15	15	6	12	10
Y	15	12	16	16	24	13	20	18	16	24

TIME CONSTANTS FOR F_2 IN MILLISECONDS

TABLE 5.3e – CONSONANT – VOWEL TRANSITION MATRIX

	IY	I	E	AE	UH	A	OW	U	OO	ER
B, P, M	20	16	33	10	25	15	16	8	10	50
D, T, N	50	22	10	16	16	33	67	33	10	40
G, K, NG	10	20	16	10	16	67	50	33	40	67

TIME CONSTANTS FOR F_3 IN MILLISECONDS

TABLE 5.3f – VOWEL – CONSONANT TRANSITION MATRIX

	B P M	D T N	G K NG	F V	TH THE	S Z	SH ZH	W	L	R	Y
IY	30	16	26	20	20	12	12	30	6	32	30
I	8	56	40	20	32	12	20	40	6	38	24
E	16	33	25	20	20	32	60	50	6	40	32
AE	10	33	42	20	40	40	60	74	6	40	32
UH	13	33	16	50	40	60	66	12	6	24	48
A	15	33	67	20	50	54	70	44	6	30	26
OW	10	16	20	26	32	26	66	20	6	30	40
U	16	16	13	10	20	26	30	12	6	12	36
∞	13	10	13	20	20	12	20	12	6	24	32
ER	13	33	30	20	46	40	66	32	6	20	48

TIME CONSTANTS FOR F_1 IN MILLISECONDS

TABLE 5.3g - VOWEL-CONSONANT TRANSITION MATRIX

	B P M	D T N	G K NG	F V	TH THE	S Z	SH ZH	W	L	R	Y
IY	20	67	26	52	66	44	50	30	36	32	30
I	16	58	42	40	66	52	50	40	24	38	24
E	15	10	28	26	60	52	30	50	36	40	32
AE	30	16	40	40	60	60	30	74	42	40	32
UH	15	33	25	38	24	34	80	12	12	24	48
A	15	33	33	20	20	34	70	44	18	30	26
OW	10	33	30	20	32	26	66	20	12	30	40
U	16	33	16	10	30	26	66	12	12	12	36
∞	33	30	20	50	40	64	104	12	18	24	32
ER	10	40	30	20	46	60	66	32	30	20	48

TIME CONSTANTS FOR F_2 IN MILLISECONDS

TABLE 5.3h – VOWEL-CONSONANT TRANSITION MATRIX

	B P M	D T N	G K NG								
IY	20	50	26								
I	16	60	20								
E	33	10	16								
AE	25	16	40								
UH	25	16	16								
A	15	33	33								
OW	16	67	30								
U	8	33	33								
∞	30	10	40								
ER	50	40	67								

TIME CONSTANTS FOR F_3 IN MILLISECONDS

formant with respect to a reference time, (perhaps the time of initiation of motion of formant one).

The general case of matrices of formant delays was not implemented. Preliminary experimentation showed that delay of formant motion was necessary for only three cases.

The first case involved vowel-consonant transitions where the consonant was a stop or nasal. The point of initiation of the transition of formant one was delayed (with respect to a reference time which will be explained in a later section) by ($\tau 2$-$\tau 1$) msec where $\tau 2$ is the time constant for formant two and $\tau 1$ is the time constant for formant one. [If ($\tau 2$-$\tau 1$) is negative, the delay is made zero.] This delay served to lengthen the time for which the transitions of formants two and three (both undelayed) were of significance.

The second case involved consonant-vowel transitions for all consonants except W, L, R, and Y. An identical delay of ($\tau 2$-$\tau 1$) msec was used for formant one. This delay again served to emphasize the transitions of formants two and three. (Experiments have shown that the transitions of formants two and three are extremely important in identifying consonants.)

The third case involved transitions of D and T to all phonemes. Here the point of initiation of formant two's transition was delayed until the time of release of the stop, or until the source amplitudes were switched on. The purpose of this delay was to emphasize the starting value of 1700 Hz for formant two transitions for D and T.

5.4.6 Timing of Formant Changes

One of the major features of the synthesis strategy is the decision algorithm for initiating motion (of formants) to a new phoneme. There must

be some criterion to decide when formants should begin moving towards targets appropriate for a new phoneme. This criterion can be based on a variety of decisions. It can be based on an overall duration check for each phoneme. When the duration of transition plus steady state has exceeded some value specified for the phoneme, the decision to proceed to the next phoneme can be made. This criterion is unreasonable, as we have previously argued, since phoneme duration need not be and is not invariant to context.

The criterion can be based on a duration of steady state alone. This is also unreasonable for reasons identical to those above.

The decision to initiate motion to a new phoneme can be based on the criterion that the formants must first be within the frequency regions of the targets, and then satisfy durational requirements of the phoneme, if there are any. This criterion seems reasonable in that the majority of the time phoneme durations will be unspecified and will vary with context. Only in a few specific cases will constraints, external to the internal workings of the strategy, be used.

There may be other criteria but from the point of view of the rest of the scheme for synthesis, these few seemed to be the most reasonable ones to choose from.

Thus our criterion for initiating formant transitions is based on the frequency regions, context, and external constraints on duration.

We are now in a position to discuss the general case of formant motion, and concomitantly the internal mechanisms for control of timing.

Formants, in general, are in motion towards target values appropriate for the phonemes to be generated. Their motion is characterized by the solution to a differential equation. The time constant of motion is a

function of the phoneme from which motion began and the phoneme which is being generated. Each formant moves with its own time constant and there is provision for delay in time of initiation of the motion of formants.

The major problem concerns how we determine when to stop motions towards one target, and proceed to another. As emphasized previously, this is the function of the frequency regions around the targets. When all formants are within the frequency regions of the target a decision is made. If a stressed vowel is being generated then a table lookup procedure determines the correct vowel duration and motion continues for the specified time duration. Once a vowel of proper duration has been generated motion towards target positions characteristic of the next phoneme is initiated. If the current phoneme is not a stressed vowel, motion towards the new phoneme targets is initiated as soon as all the formants are within their specified frequency regions.

The initiation of motion to new targets is manifested in three ways. First new time constants for each formant are inserted into the respective difference equations. Second, the forcing functions (input) to the difference equations are changed in a step-like manner indicating the changes in target positions. Finally the initial conditions of the difference equation are set to preserve continuity of formant values and formant velocities. If any formant motion is to be delayed, the changes in the difference equation for that formant are delayed appropriately.

The example of Fig. 5.4 should help to clarify the above descriptions. Initially formants one and two (we shall neglect formant three in this example) are at target positions appropriate for phoneme 1. At time t1 motion is initiated to phoneme 2. Formants one and two begin motion

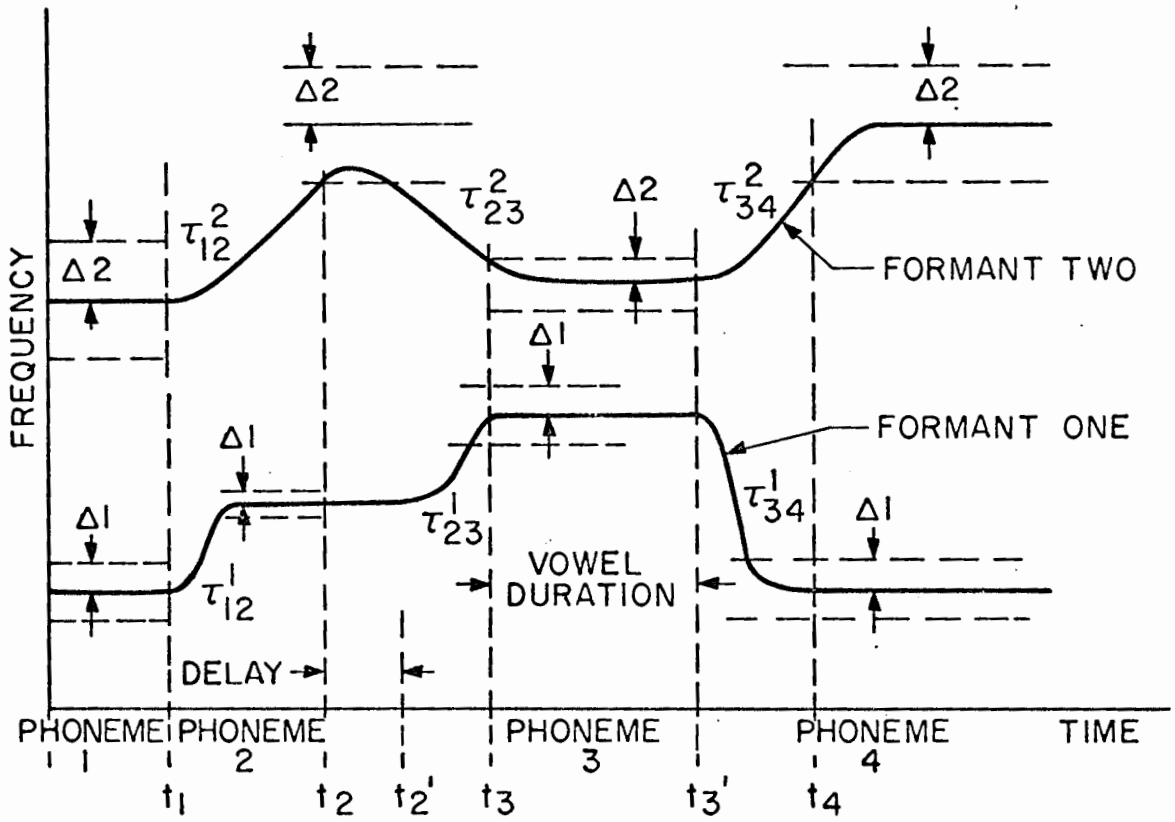


FIGURE 5.4

SIMPLIFIED EXAMPLE OF FORMANT MOTION

simultaneously (no delay is used here) with time constants τ^1_{12} and τ^2_{12} respectively. τ^1_{12} is much smaller than τ^2_{12} so formant one moves more rapidly to its target value than formant two. Periodically the formant values are tested to see whether or not they are within the specified frequency regions of the targets. (The frequency regions are indicated by $\Delta 1$, $\Delta 2$ in Fig. 5.4.) If they are not the formants continue their motion thus moving closer to target. For the example in Fig. 5.4, formant one enters its frequency region prior to formant two. Until formant two enters its frequency region at time t2, formant one moves closer to its target position. At time t2 both formants are within the specified frequency regions and so a check is made on whether phoneme 2 is a stressed vowel or not. In this example phoneme 2 is not a stressed vowel, so motion to phoneme 3 is initiated at t2. However we now have the case when the time of initiation of motion of formant one is delayed. Hence at t2 the target value and time constant for formant two is changed but formant one's target is unchanged. At time $t'2$ the delay is terminated and formant one initiates its motion.

Phoneme 3 of Fig. 5.4 is a stressed vowel. So at time t3, when both formants are within the frequency regions for phoneme 3, motion to targets for phoneme 3 continues for the specified vowel duration. At time $t'3$, following the vowel duration, motion is initiated towards targets for phoneme 4. New time constants and targets are again inserted in the equations of motion. This process continues in this manner until all phonemes have been generated.

The only external constraints on timing seen from the example of Fig. 5.4 are those on vowel duration. However there are additional constraints on maximum durations etc. We shall see how these constraints

affect the above simplified example once we understand how the remaining synthesizer parameters are generated.

5.4.7 Timing of Source Amplitude Changes

Until now we have been primarily discussing the dynamic aspects of generating formant data. Clearly the formant data, as we are using it, is of greatest importance in that the basic timing mechanisms are based primarily on the formants. However there are nineteen synthesizer parameters and the formants only comprise six of them (three bandwidths and three center frequencies). The remaining thirteen parameters can be broken up into three groups. One group contains source and gating amplitudes (A_N, A_V, A_{VB}, THRESH), one group contains nasal and fricative poles and zeros, and the last contains fundamental frequency. In this section we will be concerned with the first group.

We have previously discussed both the rates at which source parameters must switch, and the levels to which they must head. Now we must decide when the switching should begin. The switching should occur some time after initiating motion (formant changes) to a new phoneme and well before initiating motion to the succeeding phoneme. The longer you wait after initiating motion to switch source characteristics, the greater amount of the transition is lost (or preserved). For example for transitions from a stop to a vowel, the longer you wait to switch, the less of a transition you get (since voicing amplitude is zero before the transition) and the more difficult it is to identify the stop consonant. For transitions from vowels to stops, on the other hand, the longer you wait to switch, the more of a transition you get and the easier it is to identify the stop consonant. This simple example should point out that there is no single time at which

switching should occur but instead it is a function of context. A compromise value can be chosen but suitable modifications must often be made.

As seen above, the times at which source amplitudes change are time-locked to the formant motion. If we consider the case of formants proceeding from steady state values to new steady state values, then the transition time is about 4τ MAX where τ MAX is the largest of the formant time constants. The slope of a formant transitions is maximum at a time τ msec after the transition is initiated where τ is the time constant. At time τ msec after the transition is initiated about 30% of the transition has taken place. τ msec later, about 65% of the transition has occurred. For consonant-vowel transitions, we generally desire a large percentage of the transition to occur after the source characteristics have changed (i.e. with the voice source turned on) and thus the switching takes place at a time τ_1 msec after the initiation of formant motion where τ_1 is the time constant for formant one. (The delay of initiation of the motion of formant one was shown to be $\delta = \tau_2 - \tau_1$ msec where τ_2 is the time constant for formant two and τ_1 the time constant for formant one. Hence a time τ_1 msec after initiating formant one motion will correspond to a time $\tau_1 + \delta = \tau_2$ msec after initiating formant two motion. At this time 30% of the transitions for both formants one and two will have taken place.)

For vowel-vowel transitions there are no changes in source characteristics (the voicing amplitudes of all vowels being equal) and hence we need not worry about this case.

For vowel-consonant transitions we again change source characteristics at a time τ_1 msec after initiating formant motion. (Again we have set the delays so 30% of the transitions for both formant one and

formant two will have occurred at this time.) However for the case when the consonant was a stop, turning off the voice source after only 30% of the transition of formant two had taken place proved inadequate. Experiments with the synthetic utterances (see the next chapter for a thorough discussion) showed that more of the transition had to be voiced. Hence the time when the sources were switched was changed to a compromise value of 1.5τ msec after the transition had begun at which time about 45% of the transition had occurred.

For consonant-consonant transitions similar results were found. When the second consonant was a stop, the source characteristics were switched at a time 1.5τ msec (τ again is the time constant of formant one) following the initiation of motion in order to preserve a large percentage of the transition (as from a nasal or a voiced fricative). Otherwise the source characteristics were switched at a time τ msec following initiation of formant motion.

The slope of the formant transitions is generally maximum at a time τ msec after they begin, and monotonically decreases thereafter. Hence for any time greater than τ msec after the transitions have begun, the initial slope of the formant transitions, after the sources have been switched, is always steepest. When one examines spectrograms of phoneme transitions (B $\rightarrow$ IY for example) the initial observed slopes of the transitions are usually greatest and they indeed seem to be monotonically decreasing as the transition progresses.

The assumption that the transition begins from a steady state and hence that the maximum slope occurs τ msec later and the transition takes 5τ msec is not valid in general. For the case when there is an

initial formant velocity these numbers are modified somewhat. Maximum slope of a formant transition from Ai to Af with initial velocity Vi occurs at:

$$\tau' = \tau \left[1 + \frac{Vi\tau}{Vi\tau+(Ai-Af)}\right] \qquad (5.11)$$

Generally the variation in time of the point of maximum slope can be neglected. [This is true when (Ai-Af) in Eq. (5.11) is much greater than $Vi\tau$.]

For nasal consonants formant one is generally discontinuous at the onset of the nasal. This discontinuity is due to the sudden closing of the oral passage and hence a shift of formant one from being the lowest resonance of the oral-coupled vocal tract. This discontinuity occurs at the onset of the nasal, i.e. the time when the source amplitudes switch. (For nasals the voicing amplitude is somewhat smaller than for vowels.)

5.4.8 Timing of Shifts of Nasal and Fricative Poles and Zeros

For the production of nasals and fricatives, their respective poles and zeros must be properly positioned at the time the source characteristics switch. Hence they must begin moving away from their rest positions prior to the source switching time.

The nasal pole and zero are known to be present generally 50-100 msec prior to the nasal onset. This is due to the relatively long time it takes to lower the velum. When a vowel precedes the nasal we get a nasalized vowel. Hence the nasal pole and zero can initiate motion long before the nasal onset and for vowels preceding the nasal

this would be highly desirable. In other cases, the nasal pole and zero could initiate motion at the time that formants initiated motion.

The fricative pole and zero cannot begin motion too soon as this does not occur physiologically (as with prenasalization of vowels). Hence the earliest time the fricative pole and zero could initiate motion is the time the formants begin moving to new targets.

In this implementation, the time at which we initiate formant changes is also the time when nasal and fricative poles and zeros initiate motion. This motion is linear and the slope is so arranged that the poles and zeros just reach their targets at the time the source amplitudes are switched.

For nonnasal sounds, the target positions of the nasal zero and pole are set to 1400 Hz. Thus the pole and zero will cancel each other in these cases. Further, for nasal sounds, the bandwidth of formant one (nominally 60 Hz) is changed linearly to 150 Hz for the duration of the nasal. The bandwidth begins to change at a time 50 msec prior to the time at which the amplitudes switch and is linearly changed back to its nominal value in 50 msec after the nasal.

For nonfricative sounds the fricative pole and zero target positions are set to 1500 Hz.

The motion of the poles and zeros has been chosen to be linear. For fricatives (except for the case when two fricatives follow each other) it makes no difference what type of motion we use as the noise source is not turned on until the pole and zero are properly positioned. For nasal sounds the motion of the pole and zero will be of significance for cases when the nasal is preceded or followed by a voiced sound. For these cases

there is no clear evidence for or against linear motion and for simplicity it was used.

5.4.9 Contextual Effects

Provision must be made, in the general framework, for changes in phoneme characterization as a function of context. Certain stop consonants, G for example, are known to have shifting points of articulation, depending on the succeeding vowel. It is expected that the target positions of the formants would similarly change for G as a function of the succeeding vowel.

The following changes in phoneme characterizations have been found necessary and are included in this implementation.

1. The formant target positions for R in initial word position are changed.
2. The target position of formant two for B, P, and M preceded by the back vowels OW and OO is set to 1300 Hz from its normal value of 800 Hz.
3. The target position of formant two for F and V preceded by the vowel OO is set to 1300 Hz.
4. The target position of formant two for TH and THE preceded by the vowel OO is set to 1800 Hz.
5. The target position of formant two for S and Z preceded by a back vowel is 1800 Hz.

Many of these contextual effects were determined from tests (described in Chapter Six) and therefore do not always reflect the most general cases.

5.4.10 Fundamental Frequency Contour From Features

Physiologically, fundamental frequency is controlled mainly by the tension on the vocal cords of the larynx (laryngeal tension) and the air

pressure just below the glottis (subglottal pressure). (Laryngeal tension is a representation of many effects which control the fundamental frequency of the vocal cords when subglottal pressure is held constant.)
An increase in subglottal pressure (Ps) with laryngeal tension (LT) held constant will cause the fundamental frequency (FO) to rise. An increase in laryngeal tension, with subglottal pressure held fixed, will also cause fundamental frequency to rise. Hence an increase in fundamental frequency may be reflected either in increased subglottal pressure or increased laryngeal tension.

Our model for fundamental frequency generation is based on the assumption that fundamental frequency can be derived from laryngeal tension and subglottal pressure. The limited amount of data available tends to confirm this assumption. (A problem here is that laryngeal tension is not easily measured and in general must be deduced from other data. Clearly our assumption about deriving FO from Ps and LT is only as good as the deductions about LT.) Accepting our assumption as valid, we must arrive at a description of Ps and LT that is valid for an utterance, and is modifiable as a function of context. Hence some description (articulatory, acoustic etc.) of the properties of an utterance, in terms of our variables, must be used. The work of Lieberman (1966) appears to have the most relevance here.

Lieberman made a linguistic analysis of American English in terms of the features and rules underlying the observable phenomena. A phonetic feature of importance, according to Lieberman, was called the breath-group. As stated by Lieberman, an unmarked breath-group would be characteristic of a simple, declarative sentence of American English. Lieberman said,

"The breath-group at the articulatory level involves a coordinated pattern of muscular activity that includes the subglottal, laryngeal, and the supra-glottal muscles during an entire expiration. The data shows that the sub-glottal respiratory muscles start to force air out from the lungs while the laryngeal muscles close the glottis to its 'phonation neutral position' and adjust the tension of the vocal cords. The supraglottal vocal tract simul-taneously begins to move into position and phonation commences at a specified fundamental frequency that the speaker can repeat at will. At the end of the sentence the subglottal respiratory muscles lower the subglottal air pressure during the last 150-200 msec of phonation. The tension of the laryngeal muscles for the unmarked American English breath-group appears to remain relatively steady throughout the sentence." The fundamental frequency of phonation is thus a function of the subglottal air pressure function and it falls during the last 150-200 msec of phonation. This unmarked breath-group, a suprasegmental feature, would be modified as a function of two variables - laryngeal tension, LT, and subglottal pressure, Ps. LT would remain fixed throughout an unmarked breath-group. The last 150-200 msec of phonation, for a question or marked breath-group, would be characterized by an increase in LT. Normally, Ps is falling at the end of a breath-group. Thus, an increased LT for a marked breath-group would counter the falling Ps, causing F0 to have a "not falling" terminal contour.

The problem of generating fundamental frequency, using the above concepts, is reduced to the problem of characterizing laryngeal tension and subglottal pressure for an unmarked breath-group and making suitable modifications for context. The archetypal LT contour for the unmarked

breath-group, as suggested above, is constant. The archetypal subglottal pressure contour has also been suggested by Lieberman's data. Since LT is constant throughout most of the breath-group, fundamental frequency variation is attributed entirely to subglottal pressure. A plot of this archetypal Ps contour is shown in the solid curve of Fig. 5.5. Subglottal pressure increases over the first 300 msec of the breath-group and then stays constant until the last 300 msec of the breath-group.

To complete our picture we must include the effects of the phonemes on subglottal pressure, and the marked breath-group on laryngeal tension. The effects of a marked breath-group on LT are clear. During the last 175 msec of phonation, the increase in LT must compensate for the decrease in Ps to give fundamental frequency (which is linearly proportional to the sum of Ps and LT) a rising terminal contour. A slope of 0.6 Hz/msec was assigned to LT for the unmarked breath-group. This generally provided a terminal rise of 60 Hz.

The subglottal pressure contour was modified by both consonants and vowels. Two levels of stressed vowels were adopted. One level was referred to as emphasis and only one vowel in a breath-group could be emphasized. The emphasized vowel provided the highest peak in the Ps contour. All other stressed vowels were treated similarly. When a vowel was stressed, there was an increase in subglottal pressure for a period of 500 msec, centered on the vowel. An example of this effect is seen in Fig. 5.5. The dashed curve shows the effects of placing stress on a vowel at the beginning of the breath-group. There is a rise in Ps early in the breath-group and the increase is centered at ts, the midpoint of

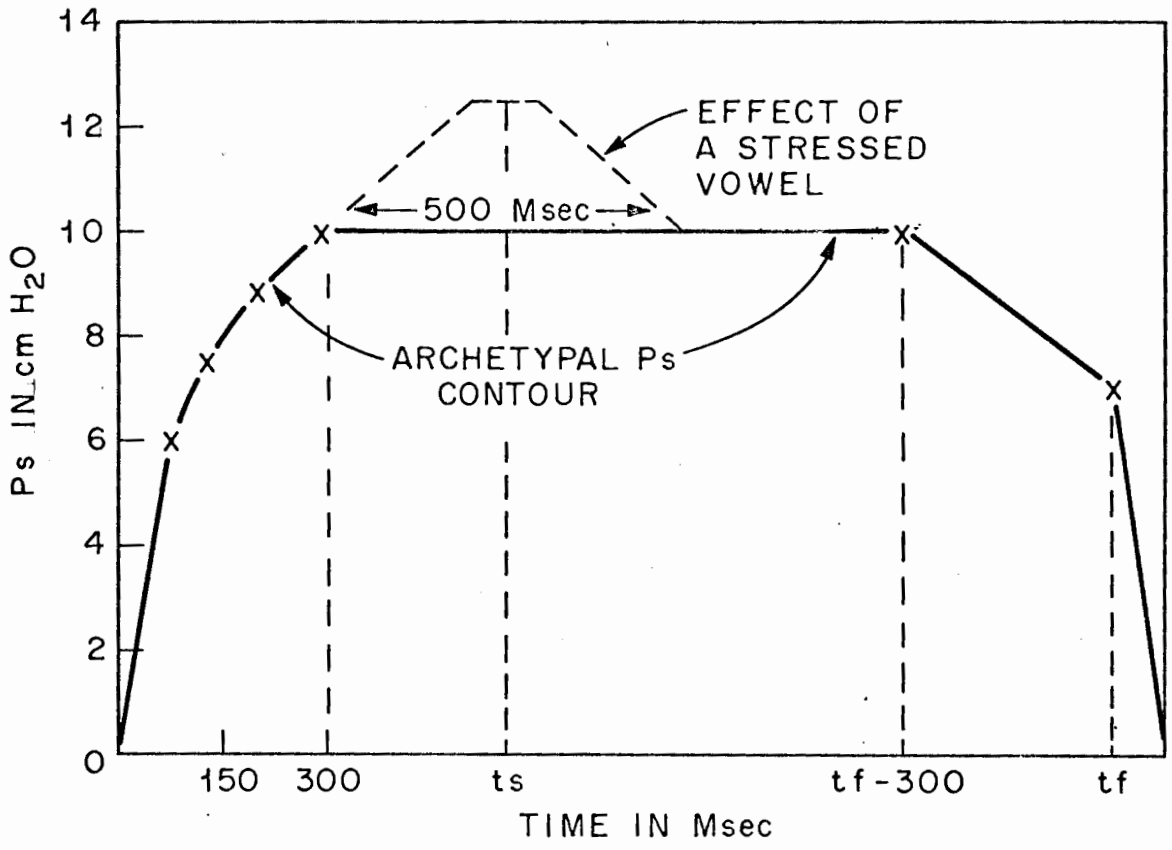


FIG. 5.5

ARCHETYPAL SUBGLOTTAL PRESSURE CONTOUR SHOWING EFFECTS OF VOWEL STRESS

the vowel steady state. If the stressed vowel had been the emphasized one, the only difference in Fig. 5.5 would be the amplitude of the increased Ps. It would have been 2.5 cm H_2O as compared to 1.0 cm H_2O.

The effects of consonants on subglottal pressure have also been included in our scheme. For a voiceless consonant, subglottal pressure automatically increases whereas subglottal pressure automatically decreases for voiced consonants. Thus the consonants introduce local perturbations to the Ps contour, centered on the consonant midpoint. The change in Ps for consonants is ±1.0 cm H_2O and this change occurs over a period of 150 msec centered on the consonant.

Figure 5.6 shows a typical fundamental frequency contour for the synthetic utterance AE-T-UH where the vowel AE is emphasized. The effects of the T are seen in the region where F0 exceeds 180 Hz. As indicated previously, fundamental frequency is linearly related to the sum of laryngeal tension and subglottal pressure.

This method for generating fundamental frequency contours has a firm basis in the physiology of how F0 is controlled. The rules for changing Ps and LT are simple and straight-forward and based on easily observable phenomena. This procedure is one of the major points of this thesis work. The extent to which it has been successful in accomplishing its aims will be discussed in Chapter 7.

5.5 Examples of Input Strings

The input for this synthesis strategy contains information as to phonemes, pauses, word and sentence boundaries and vowel stress. At this point it would be of value to see the exact form of the input.

Table 5.4 gives the computer input for ten sentences. All vowels are unstressed except those followed by the symbol STRSS. The symbol

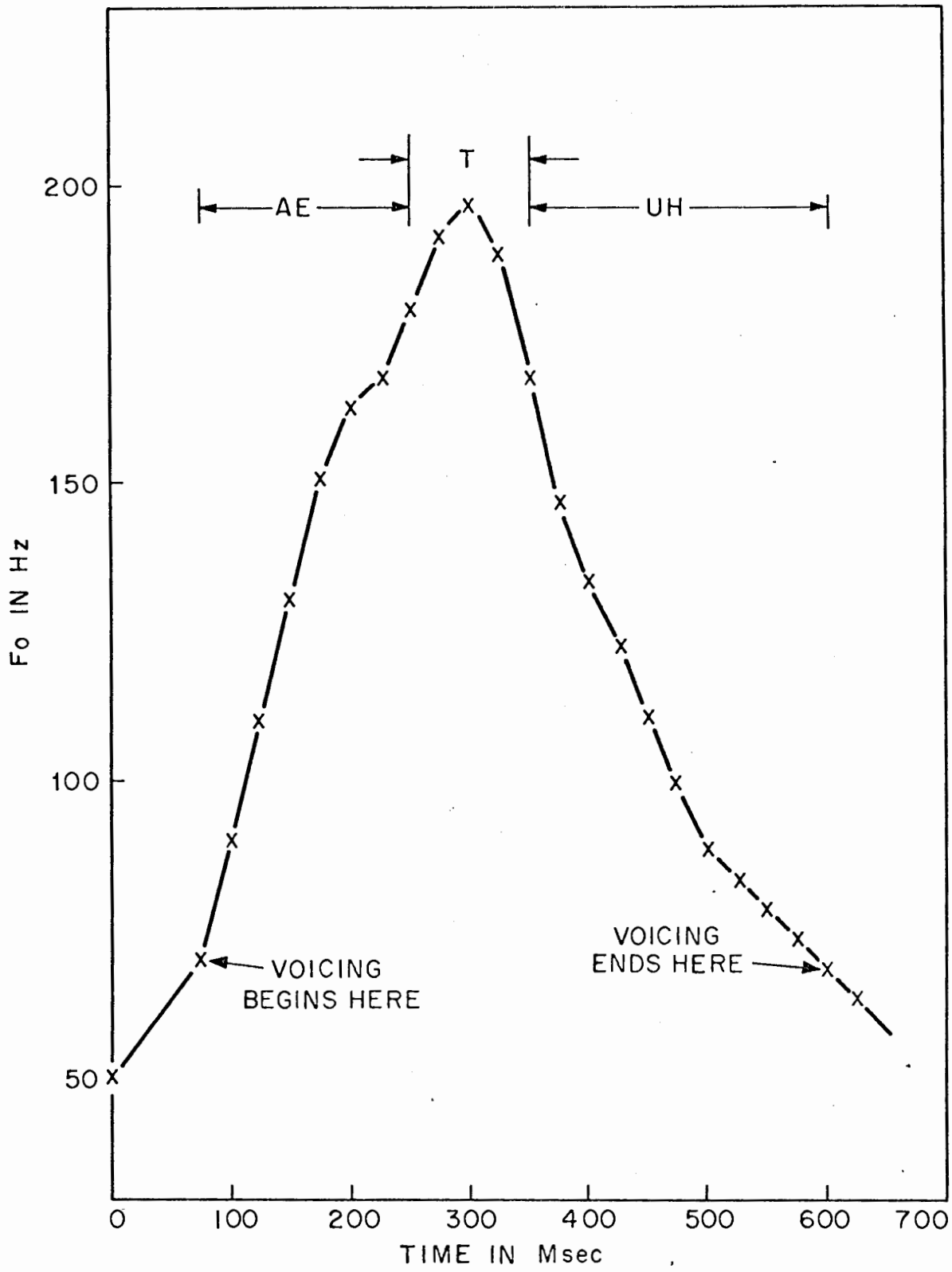


FIG. 5.6

EXAMPLE OF FUNDAMENTAL FREQUENCY CONTOUR FOR SYNTHETIC SYLLABLE - AE-T-UH

TABLE 5.4

EXAMPLES OF TYPICAL INPUT STRINGS FOR SYNTHESIS STRATEGY

1. This is an olive.

   THE I STRSS S SPACE I Z SPACE AE N SPACE A STRSS1 L I V END

2. Why are you sad?

   W A STRSS1 IY SPACE A R SPACE Y OO SPACE S AE STRSS D END

3. We sang all day.

   W IY SPACE S AE STRSS1 NG SPACE OW L SPACE D E IY STRSS END

4. Give me the book.

   G I STRSS1 V SPACE M IY SPACE THE UH SPACE B U STRSS K END

5. Are you feeling well?

   A STRSS R SPACE Y OO SPACE F IY STRSS1 L I NG SPACE W E L QUES END

6. Don't eat the soup.

   D OW OO STRSS N T SPACE IY STRSS1 T SPACE THE UH SPACE S OO STRSS P END

7. They were all well.

   THE E IY SPACE W ER SPACE OW STRSS1 L SPACE W E STRSS L END

8. Larry and Bob are here.

   L AE STRSS R IY SPACE AE N D SPACE B A STRSS1 B SPACE A R SPACE

   H I STRSS R END

9. We heard a loud noise.

   W IY SPACE H ER STRSS D SPACE UH L A STRSS1 OO D SPACE N OW STRSS .

   IY Z END

10. The sound was unclear.

    THE UH SPACE S A STRSS1 OO N D SPACE W A Z SPACE UH N K L IY STRSS R END

STRSS1 signifies the emphasized vowel. Word boundaries are signified by the symbol SPACE and pauses by the symbol PAUSE. A question is signified by the symbol QUES. The sentence boundary is indicated by the symbol END.

5.6 Further Aspects of the Strategy

In this section we will discuss three further aspects of this implementation. The first of these concerns use of the input symbol SPACE which signals a word boundary. Whenever a word boundary occurs, consonants on either side of the word boundary are affected. In this strategy an initial consonant is lengthened by about 20% whereas a final consonant is shortened by a similar amount. A word boundary has no effect on phonemes which do not lie on either side of the word boundary. Undoubtedly this is a shortcoming of our implementation.

Voiceless stop consonants following S are not aspirated in general. However if a word boundary follows the S then the voiceless stop is aspirated. This effect is included in our implementation.

The second point concerns the initialization and ending procedure for the scheme. Initialization, in general, poses no problems. For utterances beginning with vowels, a fixed steady state of 80 msec is generated and then transitions can begin to the next phoneme. (For stressed vowels the steady state is increased according to our normal rules.) For utterances beginning with nasals or voiceless fricatives a similar steady state duration of 80 msec is used. For stops, however, the scheme begins with the stop release and proceeds normally from then on. For voiced fricatives and W, L, R, and Y a steady state of 50 msec is used. The procedure for terminating the utterance is similar to that for beginning. As soon as the formants are within the frequency regions for the last phoneme a fixed

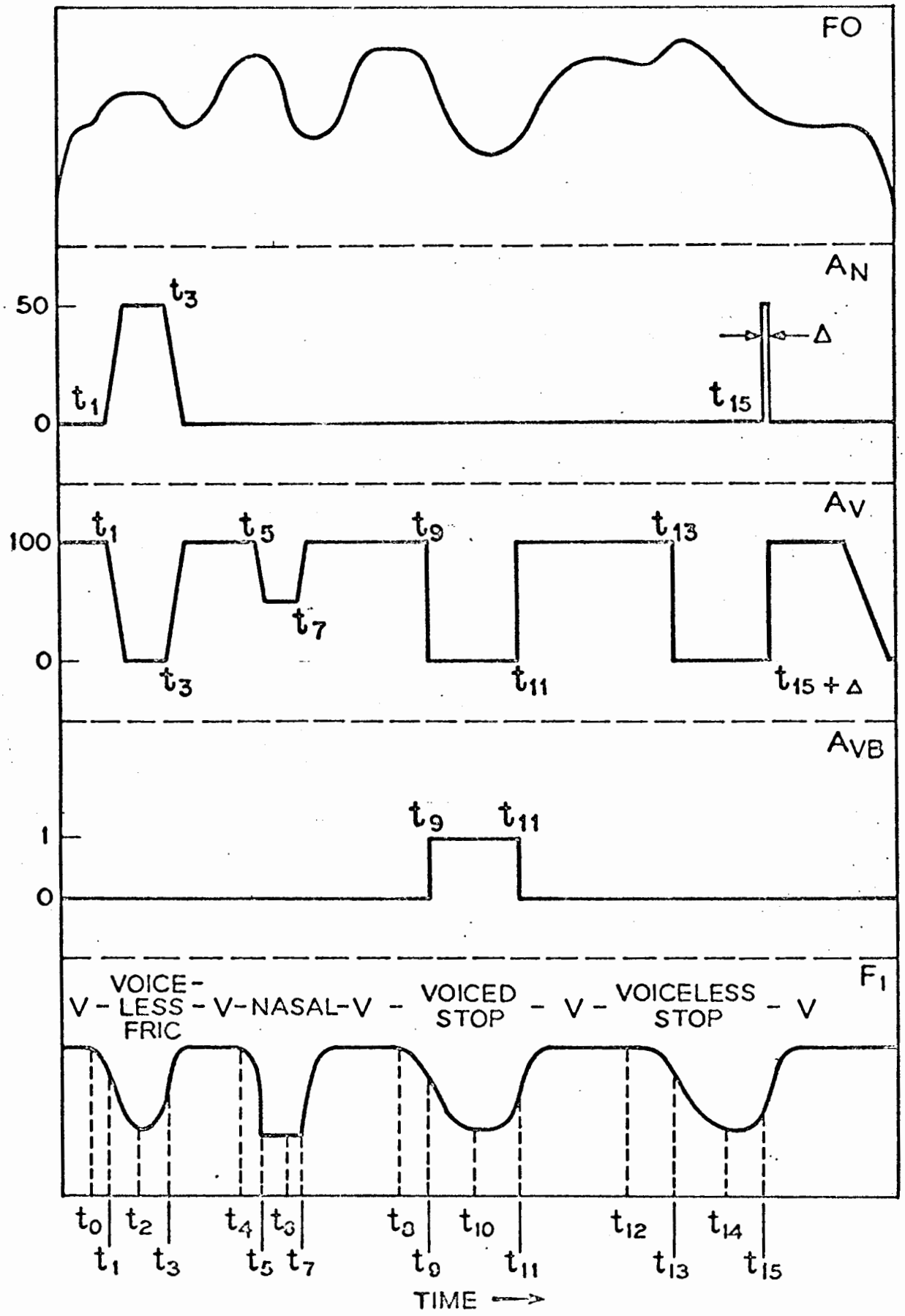
steady state duration is generated and then all sources turn off. For the voiced fricatives Z and ZH the voice source turns off much faster than the noise source hence they sound like their voiceless counterparts S and SH for the last 50 msec of the utterance.

The final point concerns the use of pauses to generate periods of silence. When a pause followed phoneme X, phoneme X was treated as though it was the final phoneme in an utterance (as described above). Following phoneme X, a pause (or periods of silence) of external specified duration was generated. Following the pause, the next phoneme was treated as though it was the initial phoneme of a new utterance. These pauses can be thought of as acting as punctuation marks. They are not generated by rule, but instead included in our input string.

Suitable length pauses need to be inserted at various points throughout an utterance. Without these pauses, a clearly intelligible sentence will often become run-on and difficult to understand. This need for pauses is a reflection of the difficulties in producing an utterance with natural timing. Errors in vowel duration, particularly gross errors, must be compensated by pauses which enable the listener to figure out what was intended. Without these pauses listeners often get confused and are unable to understand the later parts of the utterance as a result. The insertion of pauses at syntactic breaks of an utterance is a useful artifice, although difficult to do by rule.

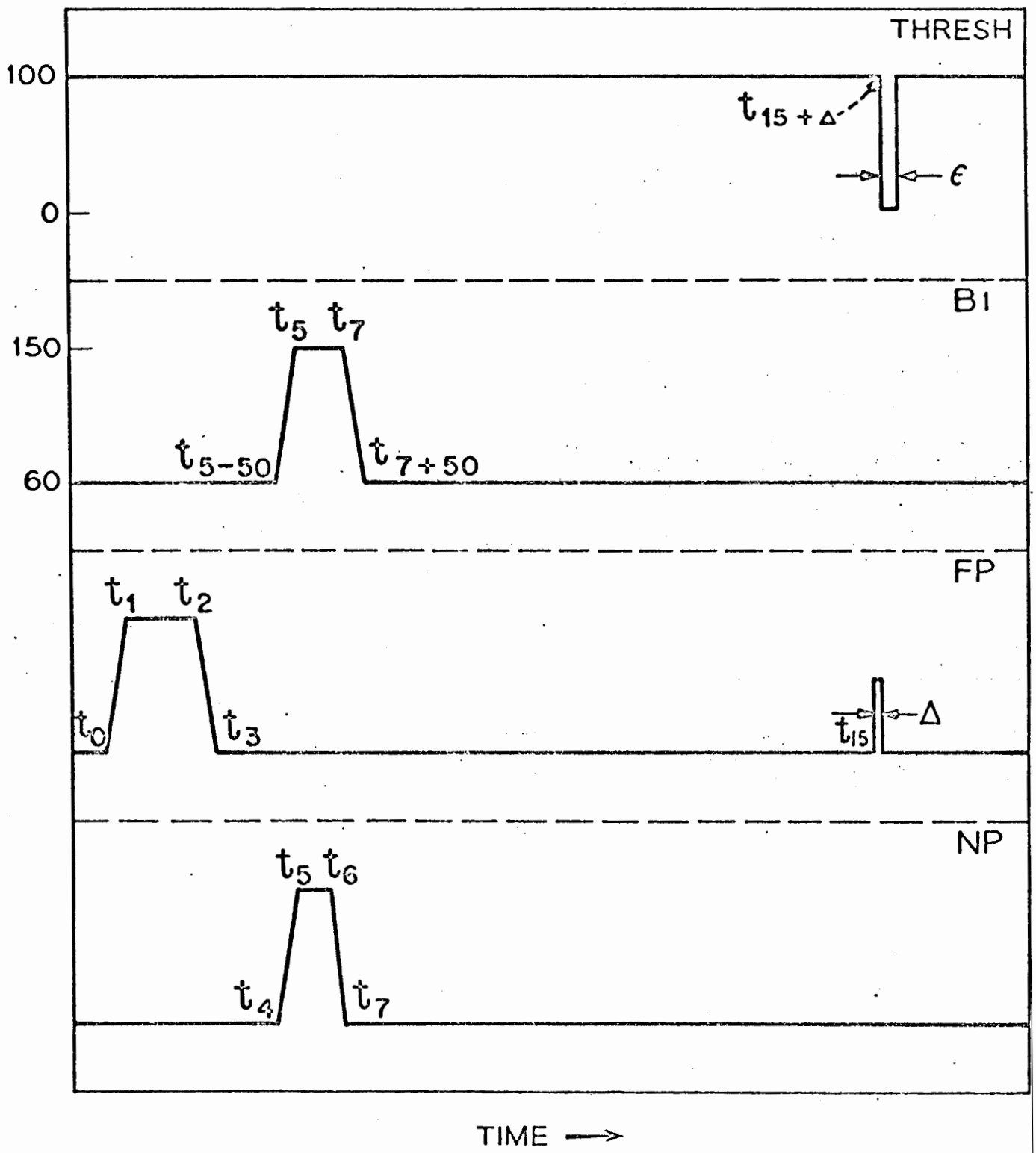
5.7 Example of Parameter Generation

Figures 5.7a and 5.7b show plots of control signals versus time generated by the synthesis strategy. The input string chosen for this example is:



TYPICAL EXAMPLE OF CONTROL PARAMETER CONTOURS

FIG.5.7A



TYPICAL EXAMPLE OF CONTROL PARAMETER CONTOURS

FIG. 5.7B

V (Vowel)-Voiceless Fricative-V (Stressed Vowel)-Nasal-V (Stressed Vowel)-Voiced Stop-V (Stressed Vowel)-Voiceless Stop-V-END

For this illustrative example we have chosen nine of the nineteen control parameters and plotted their variation during this particular input sequence.

The motion of all the control parameters is time-locked to the formant motion (in this case to the motion of formant one). Initially all control parameters are set to values characteristic of the first vowel. The noise source is off ($A_N = 0$) and the voice source is on ($A_V = 100$). At time t_0 the transition to the voiceless fricative is initiated. The proper target value and time constant is inserted into the equation of motion for formant one (F_1). At time t_0 the fricative pole (FP) initiates motion to the target value appropriate for the voiceless fricative. All other controls (except F0) remain fixed during the first transition. We shall defer the treatment of the fundamental frequency contour until the end of this section.

At time t_1, one time constant after initiation of the transition of F_1, 30% of the formant transition is completed. At this time FP is properly positioned for the voiceless fricative and remains constant throughout the fricative. The source amplitudes begin to switch at time t_1. Target values and rates of change of A_V and A_N are determined from Table 5.2. Since a voiceless sound is being generated A_V heads for zero and A_N heads for a specified value (50 in this example).

At time t_2 formant one is within the frequency region of the voiceless fricative. Since a voiceless fricative has no durational constraint the transition to the next phoneme is initiated at t_2. Again a new target position and time constant is inserted into the equation of motion for F_1. FP initiates motion back to the rest position at this time.

At time t_3, 30% of the formant transition is completed. FP is back at its rest position at this time. Hence source characteristics begin to switch to values appropriate for the vowel. (A_N heads for 0 and A_V for 100.)

Since the current vowel is a stressed vowel the initiation of formant motion to the nasal begins at time t_4, following a duration of vowel steady state. At this time the nasal pole (NP) begins motion to the proper position for the nasal.

At time t_5, 30% of the formant transition to the nasal is completed. At this time the value of F_1 sent to the synthesizer (but not the value used by the program) is discontinuous. NP is properly positioned for the nasal at time t_5. The bandwidth of formant one (B_1) was increased from its nominal value (60 Hz) to the nasal value (150 Hz) during a 50 msec interval prior to t_5. At time t_5, A_V heads for the value appropriate for the nasal.

At time t_6, the value of F_1 generated by the program is within the frequency region of the nasal so the transition to the next vowel is initiated. The target position and time constant for F_1 are changed. NP heads back for its rest position.

The synthesizer value of F_1 is held constant until t_7, the time at which 30% of the first formant transition of the program F_1 is completed. At this time the synthesizer value of F_1 is made identical to the program value of F_1. B_1 begins to head back for its nominal value at time t_7 and it reaches it 50 msec after t_7. NP is back at its rest position at time t_7. A_V begins to head for the value 100, appropriate for the next vowel, at time t_7.

At time t_8, following a duration of steady state for the stressed vowel, the transition to the voiced stop is initiated. At t_9, when 45%

of the formant transition is completed, A_V is set to 0 instantaneously. The voicebar level (A_{VB}) is set to 1 at this time and remains at this value.

At time t_{10} the transition to the next vowel is initiated. At time t_{11}, when 30% of the transition is completed, A_V is set back to 100 and A_{VB} back to 0 instantaneously.

At time t_{12} the transition to the voiceless stop begins. At t_{13}, when 45% of the transition is completed, A_V again is set to 0 instantaneously. At time t_{14} the transition to the last vowel is initiated. At time t_{15}, when 30% of the transition to the vowel is completed, two events occur. A_N is instantaneously increased and FP is instantaneously changed. (This is the initiation of the frication phase for voiceless stops.) Values are held constant for Δ msec at which time they instantaneously switch back to their original values.

At time $t_{15} + \Delta$, A_V switches instantaneously and the gating control (THRESH) switches to 0 instantaneously. (This is the initiation of the aspiration phase.) These values are held fixed for ε msec. Then THRESH switches back to 100.

After a sufficiently long vowel steady state has been generated, A_V switches to 0 and the utterance is ended.

The fundamental frequency contour shows peaks due to the stressed vowels and voiceless consonants, and valleys due to the voiced consonants. At the beginning of the utterance F0 rises from a low value and at the end of the utterance F0 falls back to a low value.

CHAPTER SIX

RESULTS AND EVALUATION

6.1 Introduction

Many of the results and much of the methodology of the previous chapters have been attained from actual experimentation - generally on an informal basis. It would be difficult to formulate all the rules for synthesizing speech without acoustic feedback enabling the researcher to test his rules. One strong merit of this method of synthesis is that it is more than an idea put down on paper. An idea is generally of unknown value until its relationship to the physical world has been established. Thus it is essential that we put our ideas to the test.

We are trying to model a physical process - the act of speaking. The measure of our success is how well both naive as well as trained listeners can understand the speech which we generate. A quantitative measure, whenever possible, has to be established in order to evaluate our method.

In this chapter we will present the results of tests of consonant intelligibility in pre-stressed and post-stressed position. Results of intelligibility tests using simple and complex sentences as test material will also be presented.

6.2 Preliminary Results

In the early months of our research, no quantitative measures were necessary. The speech was clearly unintelligible (for the most part) and even gross features of the speech were not being preserved. It was during these early months that our conception of the necessary requirements for an acoustic domain approach was greatly modified. During this period

informal listening proved adequate for identifying gross errors in our strategy. We were able to evaluate various synthesis techniques and determine aurally which ones improved the speech quality and intelligibility. A group of 21 sentences was used during this period. These sentences were chosen to provide an adequate test of all aspects of our synthesis scheme. Some were filled with contextual cues (i.e., proverbs and adages), whereas others were chosen from standard lists of sentences used in intelligibility tests(Beranek, 1949) and had few contextual cues.

Once the framework of our synthesis strategy had been adequately determined, our concentration of thought shifted to the details of our implementation. We became concerned with problems such as determining optimal times for switching source characteristics, levels of amplitude controls, durations etc. We were faced with a specific problem. Given that we accepted our framework as inviolate, how could we optimize parameters of the implementation to produce the best speech?

Our solution has been to define a criterion for "best speech." We have used as the criterion for "best speech" the speech having highest intelligibility. This is a reasonable approach since we can define a quantitative measure of intelligibility and ultimately use it as a basis of comparison with other schemes. Furthermore our primary objective is to have listeners understand what they hear. Their thoughts as to the naturalness of the speech are clearly of importance but this attribute is too subjective for quantification. (Consider the case of clearly intelligible speech played against a background of intense noise. Although the speech may be intelligible, its quality may be considered poor due to the objectionable background effects. There is some interplay between speech intelligibility and quality but we are neglecting it.)

6.3 Consonant Intelligibility Testing

Our primary measure of intelligibility has been a sequence of consonant perception tests. The lowest level of testing the intelligibility of our speech is from syllables containing one consonant, and one vowel. Hence two forms of consonant perception test were devised. (The assumption has been made that one need not ever run vowel perception tests with our scheme. Vowels produced on a terminal analog synthesizer are of high quality-- this being the forte of such a synthesizer. During the course of our consonant evaluation tests, listeners ofter wrote down the preceived vowel. An informal check of the data showed no errors in vowel identification occurred for any subject. Only five of our ten vowels were used during these tests.) One test was intended to test perception of consonants in initial word position. The schwa vowel UH always preceded the consonant and was used as a perceptual cue for stop consonants as it provided a basis for perceiving the stop gap. The second test was intended to test perception of consonants in final word position. The schwa vowel UH always followed the final consonant - again providing a basis for perceiving stop gap duration, bursts, and aspiration for the stop consonants.

Sixteen consonants were used in these tests. These included B, D, G, P, T, K, M, N, F, TH, S, SH, V, THE, Z, and ZH. Five vowels were used (beside the schwa vowel). These were IY, AE, A, OW, and OO. Thus for each test there were 80 possible stimuli. Twenty additional stimuli were used - ten initiating the test, and ten concluding it - thus giving a total of 100 stimuli per test. Only the middle 80 were used for evaluation purposes. The 80 stimuli that were to be used as our test stimuli were presented in a random order.

6.3.1 Test One

The first set of consonant evaluation tests were run primarily to determine a base line for the rules and parameters as they stood at the time of the test. The rules of the synthesis strategy were basically as described in Chapter Five with six major exceptions. These were:

1. Source parameters initiated changes τ_1 msec after the initiation of the transition of formant one in all cases (where τ_1 was the time constant of the transition of formant one).
2. The duration of aspiration following a voiceless stop consonant was a fixed value independent of context.
3. No contextual changes in target positions for formants were used.
4. No delay of initiation of motion of formant two, in transitions from D or T to other phonemes, was used.
5. No voicebar was used during the period of closure for voiced stops.
6. The bandwidth of formant one was not broadened during a nasal consonant.

Furthermore, there were many gross errors in the basic data of our strategy. The frequency positions of the bursts of voiceless stop consonants were incorrect. The relative rates of onset of voicing and noise amplitudes were wrong. The formant target positions for the voiced and voiceless fricatives were incorrect. There were also many errors in time constants of transition, and in frequency regions for certain phonemes.

The two types of tests that were conducted will be referred to as the UH-C-V test (for initial consonants) and the V-C-UH (for final consonants). Our first set of results is seen in a set of confusion matrices shown in Fig. 6.1. Three subjects (KNS, DHK, WLH) were used in the UH-C-V test and their data were pooled to give the results. Two of these same subjects (DHK, WLH) were used in the V-C-UH test and again their data were pooled to give the results shown in Fig. 6.1.

The overall percentage of correct answers for the UH-C-V test #1 was 42% whereas for the V-C-UH test #1 it was 51%. These low percentages indicated the large room for improvement. For the UH-C-V test the only consonants which were identified correctly a large percentage of the time (i.e., more than 75% of the time) were P, SH and ZH. For the V-C-UH the consonants which were generally perceived correctly were B, P, S, SH, Z, and ZH. Although the V-C-UH test produced better results than the UH-C-V test, the overall intelligibility scores on both tests were low. Furthermore, subjective comments about the quality of the consonants were unfavorable for the most part.

A cursory glance at the confusion matrices of Fig. 6.1 indicates that there are a large number of voiced-voiceless errors (i.e., a voiced phoneme is confused with its voiceless counterpart and vice versa).

6.3.2 Test Two

After a thorough evaluation of the results of our first tests, a number of changes of data were made and a second series of consonant evaluation tests were run. The same three subjects (KNS, DHK, WLH) were used in both these tests and again their data were pooled. (For the

ONSONANT R EIVED

CONSONANT SENT

	B	D	G	P	T	K	M	N	F	TH	S	SH	V	THE	Z	ZH	?
B	10		2	1			1						1				
D	3	1	8										2	1			
G		2	0	1	11							1					
P				13	1	1											
T				8	7												
K				2	4	9											
M	2	3					1	1					6	2			
N	1	4	5				1	1					3				
F									3	2	3		7				
TH									2	3	3		3	2	2		
S											7				8		
SH									1			14					
V	4		1				1						8	1			
THE	2		3										9	1			
Z											2		1	4	8		
ZH												1				14	

UH-C-V MATRIX
(3 SUBJECTS)

CONS ANT RECEIVL

CONSONANT SENT

	B	D	G	P	T	K	M	N	F	TH	S	SH	V	THE	Z	ZH	?
B	8		1	1													
D		2				1				3			3	1			
G			0		7	3											
P				10													
T				5	5												
K				5	3	2											
M	1		1				3		1					2			2
N	2		1				2	1	1					1			2
F									0	5	4						1
TH									1	3	6						
S											9				1		
SH												10					
V	4												6				
THE	1								1				3	5			
Z											2				8		
ZH																10	

V-C-UH MATRIX
(2 SUBJECTS)

TEST # 1

FIGURE 6.1

final test we will present both pooled and unpooled data.) An objection associated with using the same subjects is that there are learning effects associated with listening to synthetic speech. A subject learns to listen for certain cues and with prolonged exposure he will perform significantly better than the untrained listener. (A case in point is the author who is highly overtrained and whose results are much better in these tests than any of the subjects!) Our answer to this objection is twofold. First the amount of exposure in two twelve minute sessions on four occasions (the test routine we used) is not exceedingly great. There will be a certain amount of acclimatization to the synthetic speech, but hopefully it is no greater than the amount of acclimatization one would make when listening to the speech of someone with a heavy accent, or a new dialect. Second we have found it advantageous to use as subjects people who are well motivated. The synthetic speech is not real speech - it is only an approximation to it. An unmotivated listener will often react adversely to this strange stimulus. Our subjects were all researchers in speech, having great interest in and good knowledge of the aspects of synthetic speech. Hence their motivation and eagerness to participate as subjects were of great value.

The confusion matrices for our second test are shown in Fig. 6.2. An analysis of the data presented in Fig. 6.2 shows that the overall percentage correct has risen to 60% for consonants in initial position, and 55% for consonants in final position. Furthermore the number of initial consonants correctly identified 75% of the time or more rose to six. These were B, P, S, SH, Z, and ZH. In final position, the number of consonants identified correctly 75% of the time or more was five. These were P, S, SH, Z, and ZH.

CONSONANT RECEIVED

CONSONANT SENT	B	D	G	P	T	K	M	N	F	TH	S	SH	V	THE	Z	ZH	?
B	13		1	1													
D	2	2	6		1	1								2			1
G	2	1	11														1
P				12	2	1											
T				6	5	4											
K				3	6	6											
M	9		2				4										
N		4	1					5					2	3			
F									7	6	2						
TH									3	10	2						
S											15						
SH												15					
V	8					1							6				
THE		1					2						8	4			
Z											2				13		
ZH																15	

UH-C-V MATRIX
(3 SUBJECTS)

CONSONANT RECEIVED

CONSONANT SENT	B	D	G	P	T	K	M	N	F	TH	S	SH	V	THE	Z	ZH	?
B	11			2									2				
D	1	1	8			3											2
G	2	1	8	1													3
P				15													
T					1	14											
K				5	10	0											
M	8						3						2	1			1
N		3	4					4					2	2			
F									8	5	2						
TH									1	11	3						
S											15						
SH												15					
V	6												9				
THE	2												9	4			
Z											1			2	12		
ZH												1				14	

V-C-UH MATRIX
(3 SUBJECTS)

TEST #2
FIGURE 6.2

Our data showed significant improvements for many of the consonants. The number of confusions which involved errors in more than one dimension (the three dimensions are place, manner, and voicing) decreased significantly from test #1 to test #2. Furthermore, the number of unfavorable comments about the consonants decreased.

6.3.3 Test Three

After a period of spectrographic analysis of the data of test #2, a substantial number of changes were made. The major ones involved correcting the burst frequency positions for T and K; making the aspiration duration variable depending on whether the vowel was stressed or not; changing the target positions for D, T, G, and K before back vowels; raising formant two target position for V; and finally uniformly changing source amplitudes at a time 2τ msec after the point of initiation of formant motion - instead of at time τ msec after this point. This last change was included to provide information about the optimal time to switch source characteristics.

The technique of making so many changes between tests is a dangerous one. The changes are not independent ones and many consonants are affected by more than one change. Hence, it is impossible to evaluate any of the changes individually. Our explanation as to why we did this is twofold. First, the majority of the changes were necessary and had to be beneficial. Second, if we made the changes one at a time and then ran a test each time, we would have to run 62 tests to evaluate just 6 possible changes (i.e., 6 tests to evaluate single changes, plus 15 tests to evaluate pairwise changes, etc.). This is impractical. With the understanding that the results of the test would reflect a lot of changes we proceeded to test #3.

The tests were again run on the three subjects of our previous test. The confusion matrices from test #3 are shown in Fig. 6.3. (The data for the three subjects has again been pooled.)

For consonants in initial position the overall percent correct was 69, and for consonants in final position the percent correct was 65. The number of initial consonants identified correctly more than 75% of the time was seven. These were B, P, T, S, SH, Z, and ZH. The number of final consonants in this category was also seven and these were identical to the initial consonants.

The changes made from test #2 to test #3 were not independent ones so we cannot say which ones were most influential in improving the results. However, the comments of our subjects provided cues as to which consonants sounded bad or distorted. The major change in test #3 was the switching of source characteristics at a later time in the transition. For the majority of the consonants, this did not help and in some cases it hurt (as for stops into vowels where a large part of the transition was buried in silence). Hence, in most cases, the switching time was shifted back to our original values.

The overall results of test #3 indicated a strong need for improving the stops D, G, and K and the nasals M and N. All other consonants (pooling (F, TH) and (V, THE) data) were being identified correctly 50% of the time and more.

6.3.4 Test Four

In an effort to improve the nasal consonants the bandwidth of formant one was broadened from 60 Hz to 150 Hz during the nasal. The pole and zero positions for M were found to be in error and hence changed.

ONSONANT R EIVED

CONSONANT SENT	B	D	G	P	T	K	M	N	F	TH	S	SH	V	THE	Z	ZH	?
B	13		2														
D	1	8	2				1	2					1				
G	1	2	9			2	1										
P				15													
T					14	1											
K				1	7	7											
M	9		1				2	1					1	1			
N		2	1					8	2				2				
F									7	5	3						
TH									5	9	1						
S											14				1		
SH												15					
V	3						1	1					9	1			
THE	1	1						1					7	5			
Z															15		
ZH																15	

UH-C-V MATRIX
(3 SUBJECTS)

CONS ANT RECEIVE

CONSONANT SENT	B	D	G	P	T	K	M	N	F	TH	S	SH	V	THE	Z	ZH	?
B	13		1				1										
D		4	5		1			1					2	2			
G			10			1		1		1			1	1			
P				15													
T					15												
K				9	6	0											
M	6		6				0	1					2				
N		1	2					8					1	3			
F									7	2	4		2				
TH									1	8	6						
S											15						
SH												13				2	
V	3							2					10				
THE								1	1				3	10			
Z											3				12		
ZH																15	

V-C-UH MATRIX
(3 SUBJECTS)

TEST # 3
FIGURE 6.3

For the voiced stops, a voicebar was added during the stopgap. The burst for K was made longer in duration and lower in level. The burst for K was also included for G as it was thought this would improve matters. Finally, for D and T a delay was provided to formant two so that the transition to a vowel from a D or a T always began from the steady state value of 1700 Hz. (Such effects have been noted by Stevens, House, and Paul, 1966.)

The duration for fricatives was greatly increased. Certain subjects (WLH, DHK) found this increase in duration objectionable (responding S to most voiceless fricatives) so a test was generated with everything identical but shorter fricatives. (For these two subjects the voiceless fricative results are those of this second try.)

The data of test four represents our best results, and were attained using the current synthesis strategy as described in Chapters Four and Five. Since this was our last test, the confusion matrices for both individual subjects (KNS, DHK, WLH, and the author LRR), and the pooled data for the first three subjects, are shown in Figs. 6.4, 6.5, 6.6, 6.7, and 6.8.

Figure 6.4 shows the UH-C-V and V-C-UH confusion matrices for the pooled data for test #4. For initial consonants the overall percent correct is 73 and when (F, TH) and (V, THE) responses are pooled this figure is 79%. The results for final consonants are somewhat better. The unpooled percent correct is 77 and the pooled correct percentage is 81.

The number of initial consonants identified correctly more than 75% of the time was 10. These were B, D, P, T, N, TH, S, SH, Z, and ZH.

CONSONANT RECEIVED

CONSONANT SENT	B	D	G	P	T	K	M	N	F	TH	S	SH	V	THE	Z	ZH	?
B	13												2				
D		15															
G		1	3	2	6	3											
P				13	2												
T				1	14												
K				2	7	6											
M	2		4				4						5				
N								12					2	1			
F									10	2	3						
TH									2	12	1						
S											15						
SH												15					
V	6												7	2			
THE		1											8	6			
Z															15		
ZH																15	

UH-C-V MATRIX
(3 SUBJECTS)

CONSONANT RECEIVED

CONSONANT SENT	B	D	G	P	T	K	M	N	F	TH	S	SH	V	THE	Z	ZH	?
B	12												3				
D		7	8														
G			6		1	8											
P				14		1											
T				1	14												
K				1	2	12											
M	4						4	3					4				
N			6					9									
F									11	3	1						
TH										12	3						
S											15						
SH												15					
V	3						1						10	1			
THE													4	11			
Z															15		
ZH																15	

V-C-UH MATRIX
(3 SUBJECTS)

TEST #4
FIGURE 6.4

CONSONANT RECEIVED

CONSONANT SENT

	B	D	G	P	T	K	M	N	F	TH	S	SH	V	THE	Z	ZH	?
B	5																
D		5															
G		1	2	1	1												
P				4	1												
T					5												
K					3	2											
M	1		1				3										
N								5									
F									3	2							
TH										5							
S											5						
SH												5					
V	4												1				
THE		1											3	1			
Z															5		
ZH																5	

UH-C-V MATRIX

CONSONANT RECEIVED

CONSONANT SENT

	B	D	G	P	T	K	M	N	F	TH	S	SH	V	THE	Z	ZH	?
B	5																
D		0	5														
G			4			1											
P				4		1											
T					5												
K						5											
M							2	1					2				
N			1					4									
F									3	2							
TH										5							
S											5						
SH												5					
V	2												3				
THE													1	4			
Z															5		
ZH																5	

V-C-UH MATRIX

TEST # 4-KNS

FIGURE 6.5

ONSONANT R. CEIVED

CONSONANT SENT	B	D	G	P	T	K	M	N	F	TH	S	SH	V	THE	Z	ZH	?
B	5																
D		5															
G			1		3	1											
P				4	1												
T					5												
K					3	2											
M			2				0						3				
N								4					1				
F									5								
TH									1	4							
S											5						
SH												5					
V	2												2	1			
THE													2	3			
Z															5		
ZH																5	

UH-C-V MATRIX

CONSONANT RECEIVED

CONSONANT SENT	B	D	G	P	T	K	M	N	F	TH	S	SH	V	THE	Z	ZH	?
B	5																
D		2	3														
G			2		1	2											
P				5													
T					5												
K					2	3											
M	4						1										
N		5						0									
F									5								
TH										5							
S											5						
SH												5					
V	1						1						2	1			
THE														5			
Z															5		
ZH																5	

V-C-UH. MATRIX

TEST #4-DHK

FIGURE 6.6

CONSONANT RECEIVED

CONSONANT SENT	B	D	G	P	T	K	M	N	F	TH	S	SH	V	THE	Z	ZH	?
B	3												2				
D		5															
G			0	1	2	2											
P				5													
T				1	4												
K				2	1	2											
M	1		1				1						2				
N								3					1	1			
F									2		3						
TH									1	3	1						
S											5						
SH												5					
V													4	1			
THE													3	2			
Z															5		
ZH																5	

UH-C-V MATRIX

CONSONANT RECEIVED

CONSONANT SENT	B	D	G	P	T	K	M	N	F	TH	S	SH	V	THE	Z	ZH	?
B	2												3				
D		5															
G			0			5											
P				5													
T				1	4												
K				1		4											
M							1	2					2				
N								5									
F									3	1	1						
TH										2	3						
S											5						
SH												5					
V													5				
THE													3	2			
Z															5		
ZH																5	

V-C-UH MATRIX

TEST # 4-WLH
FIGURE 6.7

CONSONANT RECEIVED

CONSONANT SENT	B	D	G	P	T	K	M	N	F	TH	S	SH	V	THE	Z	ZH	?
B	5																
D		5															
G			5														
P				5													
T					5												
K					1	4											
M							5										
N								5									
F									5								
TH										5							
S											5						
SH												5					
V													5				
THE														5			
Z															5		
ZH																5	

UH-C-V MATRIX

CONSONANT RECEIVED

CONSONANT SENT	B	D	G	P	T	K	M	N	F	TH	S	SH	V	THE	Z	ZH	?
B	5																
D		5															
G		2	3														
P				5													
T					5												
K					1	4											
M							5										
N								5									
F									5								
TH										5							
S											5						
SH												5					
V													5				
THE													1	4			
Z															5		
ZH																5	

V-C-UH MATRIX

TEST # 4 - LRR

FIGURE 6.8

The final consonants correctly identified more than 75% of the time were B, P, T, K, TH, S, SH, Z, and ZH. Perhaps equally significant results are the consonants correctly identified less than 50% of the time. For final consonants this included D, G, and M. For initial consonants it included G, K, M, V, and THE.

A detailed examination of the UH-C-V matrix of Fig. 6.4 shows:

1. Voiced Stops - B and D were almost always identified correctly. The addition of the burst for G produced ten confusions with the voiceless stops. Hence all indications are that this burst should not be used for G in our scheme. If we had used the G from test #3 our percent correct on the entire test would have been 2% higher.
2. Voiceless Stops - Only K was confused often with any other consonant. It was confused with T primarily before the back vowels (OO, A, OW). The contextual change of formant two target for K (and G) before back vowels was inappropriate.
3. Nasals - N was identified correctly a high percentage of the time and confusions occurred with D and THE indicating errors in manner but not place. M was identified correctly only 26% of the time being confused with B, G, and V. Except for the G confusions, the errors are errors in manner alone. The spectrum for M (pole and zero placement) was still not correct and this led to a large number of errors.
4. Voiceless Fricatives - S and SH were always correctly identified. F and TH were confused with each other four times and with S four times. The S confusions were entirely due to one subject (WLH).

5. Voiced Fricatives - V was often confused with B (six times) but aside from this there were no major problems. There were two voiced-voiceless confusions for Z but these were due to one subject.

Our overall conclusions are that all initial consonants were well reproduced with the exception of G, K, and M. G and K need changes in formant target positions before back vowels - reflecting the changes in place of articulation in these cases. The frication burst was inappropriate for G. (This does not imply that another burst would also be inappropriate.) The spectrum for M was a poor approximation to real M spectra due to improper pole-zero placement. There were undoubtedly other factors influencing the low intelligibility scores for M. We have not been able to isolate these factors to date.

An examination of the V-C-UH matrix of Fig. 6.4 shows the following:

1. Voiced Stops - B was confused with V three times. Otherwise B was correctly identified. Final D was often confused with G (eight times) indicating a possible need for coarticulation effects for D. Final G was often confused with K (eight times) but this is due to the burst used for G.
2. Voiceless Stops - The voiceless stops were identified correctly a high percentage of the time.
3. Nasals - The nasals were still often confused with B, V, D and G. These errors of both manner and place indicate a need for much further study into nasal characterization using this synthesis strategy.

4. Voiceless Fricatives - The voiceless fricatives were correctly identified a high percentage of the time.

5. Voiced Fricatives - The only confusions were between V and THE, and V and B. The results again indicated a high level of intelligibility for the voiced fricatives.

The only final consonants that were not well reproduced were D, G, M, and N. We have discussed the nasals in initial position and this discussion applies for final position. Nasal cues must be reinforced further. The problems for D and G are reflections of improper formant two and three values during the transition - as well as the use of the burst for G.

The confusion matrices for individual subjects for test #4 (Figs. 6.4-6.8) point out many of the aspects of our previous discussion. The data of Fig. 6.8 is for the author - a highly overtrained subject - and it points out how much acclimitization is possible. There are subject biases throughout these tests and this can be seen from the individual data. Individual correct percentages (for our three subjects) vary from about 70% to 83%.

6.4 Sentence Intelligibility

Consonant evaluation tests served as a basis for evaluating our scheme in a limited environment, and gave us a quantitative way of evaluating the effects of parameter changes. However, the percentage intelligibility score for these tests is not the only measure of our success. There are no contextual cues in these consonant tests which would enable a listener to identify a consonant which he had difficulty perceiving. Thus a more suitable test had to be used in a final evaluation of this scheme.

In anticipation of the difficulties one might encounter with our scheme, two tests were used in our final evaluation. Both tests involved using sentences as test stimuli. These sentences differed in two respects. First, one group of sentences contained only simple sentences - either declarative, exclamative or interrogative - while the other group contained complex sentences consisting of two or more clauses. The second difference between the groups of sentences concerned their makeup. The first group was not designed in any special way - whereas the second group was chosen from a list of sentences used often in intelligibility tests.

The need, in our case, for two sets of sentences is perhaps made clearer from the following quote from Beranek (1949). "The intelligibility of sentences is favored to a considerable degree by the psychological factors of meaning, context, rhythm, motivated interest, etc." Our scheme does not take rhythm into account at all. (In terms of the synthesis strategy rhythm would imply a modification of the overall timing of an utterance, and the fundamental frequency contour, based on the syntactic structure of the utterance.) The first set of sentences was designed to minimize these unfavorable effects. Rhythm was of little significance due to the inherent shortness of the sentences. Furthermore, errors in stress, as embodied by improper vowel durations, etc. were not the prime determinant of overall intelligibility for these short sentences. Hence, our first set of sentences could almost be thought of as a test of single word intelligibility, although the words were put in meaningful sentences.

For the second set of sentences, no efforts were made to minimize unfavorable effects. The sentences were chosen from a well-known set of sentences (Beranek, 1949) and presented as they were written. We shall refer to this group of sentences as the "long sentence" list, and our other set as the "short sentence" list.

6.4.1 Short Sentences

The sentences used in our test are seen in Table 6.1. These sentences were synthesized and presented to a group of listeners who were asked to write down what they heard. They were told to guess whenever in doubt.

The short sentence test was presented to eight subjects. Three of these subjects were those in our consonant evaluation tests. The remaining subjects were all untrained listeners with little or no practice in listening to synthetic speech. After one presentation of each sentence, the listener was required to write down what he heard. Then the subjects were given a second chance to listen and make changes. These tests were scored on the basis of the number of words which were correctly identified (excluding only the and a as words). The results of this test are seen below.

TABLE 6.1

LISTS OF SHORT AND LONG SENTENCES

SHORT SENTENCES

1. This is an olive.
2. Why are you sad?
3. We sang all day.
4. Give me the book.
5. Are you feeling well?
6. Don't eat the soup.
7. They were all well.
8. Larry and Bob are here.
9. We hear a loud noise.
10. The sound was unclear.
11. Do you like fish?
12. It is a cold night.
13. Stop this nonsense.
14. The sun is shining.
15. Will it snow soon?
16. I like a good book.
17. I am fine now.
18. What is your name?
19. My name is Larry.
20. You are a rat.

LONG SENTENCES

1. A yacht slid around the point into the bay.
2. The two met while playing on the sand.

3. The ink stain dried on the finished page.
4. The walled town was seized without a fight.
5. The lease ran out in sixteen weeks.
6. A tame squirrel makes a nice pet.
7. The horn of the car woke the sleeping cop.
8. The pearl was worn in a thin silver ring.
9. The fruit peel was cut in thick slices.
10. The heart beat strongly and with firm strokes.
11. The boy was there when the sun rose.
12. A rod is used to catch pink salmon.
13. The source of the huge river is the clear spring.
14. Kick the ball straight and follow through.
15. Please help the woman get back to her feet.
16. A pot of tea helps to pass the evening.
17. Smokey fires lack flame and heat.
18. The soft cushion broke the man's fall.
19. The salt breeze came across from the sea.
20. The girl at the booth sold fifty bonds.

Subject	% Correct Try #1	% Correct Try #2	Sentences Correct Try #2
JLF	93	93	18
MSL	91	93	18
MEB	92	95	18
HL	91	93	16
WLH	90	99	19
KNS	88	96	19
DHK	91	91	16
PM	98	100	20

The Sentences Correct Try #2 column is an indication of the number of sentences in which each word was correctly identified after the second trial. If we allowed single errors, these numbers would be from 0 to 3 higher for our subjects. The only sentence which was consistently identified incorrectly was sentence #3 - "We sang all day". An analysis of this utterance is presented in Section 6.6.

The subjective comments of our listeners were that the short sentences were easy to understand and often were quite natural sounding. The fricatives were rather long and this produced an unnatural effect but it did not hurt the intelligibility scores. (Most of the word errors involved errors in the voiced stops.) On the demonstration tape accompanying this thesis,* we have examples of our 20 short and long sentences.

---

* A copy of this demonstration tape is available from the author.

6.4.2 Long Sentences

The long sentences were presented to seven of the eight subjects used in our short sentence test. Two versions of this long sentence test were used. The first and second versions differed primarily in one aspect. All the sentences of our second version contained pauses at syntactic breaks of the sentences, whereas only some of the sentences of our first version contained these pauses. This modification was used at the suggestion of the three subjects who heard the first version. They felt they had done better on those sentences having pauses and suggested their inclusion for all the sentences.

The same procedure was used for the long sentence test as for the short sentence test. The tests were again scored on a word basis. The results of this test are seen below. (The first three subjects listened to version 1.)*

Subject	% Correct Try #1	% Correct Try #2	Sentences Correct Try #2
MSL	74	80	13
MEB	67	75	11
HL	81	84	13
WLH	63	67	9
KS	92	95	18
DHK	92	94	19
PM	85	89	15

* Subjects MSL, MEB, and HL were presented only 19 of the 20 sentences due to a program error.

The Sentences Correct Try #2 column includes the number of sentences in which one error or less was made. Furthermore, subject DHK indicated some familiarity with five of the sentences. Subject WLH commented that he was not very alert (recall our "motivated interest" criterion) during the test and this was certainly reflected in his low intelligibility scores (as contrasted with the other three subjects who took the same test).

6.5 Evaluation of Sentence Intelligibility Tests

As expected, the intelligibility scores were uniformly high for the short sentences and variable for the long sentences. The long sentences were less intelligible than the short sentences. This is a reflection of two factors. First, the short sentences had a lot of contextual cues (i.e., The _____ was shining. _____ = sun), whereas the long sentences had much less contextual information. The second factor was the role that stress and rhythm played in the long sentences but not in the short ones. We tried to eliminate partially these effects with pauses but this was not entirely sufficient. An understanding of the acoustic correlates of stress and rhythm is mandatory to improve both the intelligibility and naturalness of the speech.

6.6 Analysis of Two Synthetic Utterances

The majority of the short sentences were identified correctly by all subjects. In many cases the subjects felt that the utterances sounded quite natural. However one of these sentences was consistently identified incorrectly. An analysis of both this sentence and one which was always identified correctly will illustrate some of the unsolved problems.

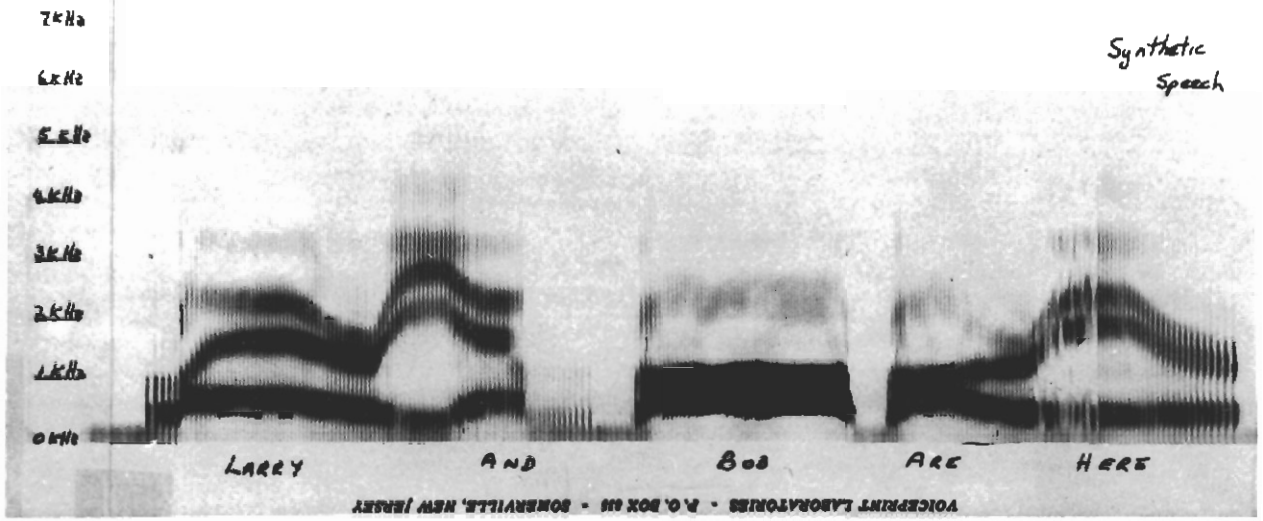
Figure 6.9 shows wideband spectrograms of the utterance "Larry and Bob are here". The spectrogram in the upper half of the figure is of the synthetic version of this utterance used in the short sentence test. The spectrogram in the lower half of the figure was made from the author's speech. (The synthetic utterance was in no way modelled after or modified by the natural utterance.) Figure 6.10 shows narrowband spectrograms of both the synthetic and natural versions of this utterance.

There is a high degree of similarity between the spectrograms of the real and synthetic speech. The durations of both the synthetic and natural utterances are comparable. The variation of the formants for both versions is quite similar as seen in Fig. 6.9. Even the fundamental frequency contours for these utterances are quite similar. As seen from Fig. 6.10, both contours are peaked during the stressed vowels A in BOB and I in HERE. A careful examination of the narrowband spectrograms shows the decrease of fundamental frequency, for both utterances, during the initial and final B of BOB.

The stressed vowels of this utterance can easily be identified from either the long steady state duration of Fig. 6.9 or the peak in the fundamental frequency contour of Fig. 6.10.

The major noticeable difference between the natural and synthetic versions is seen during the aspiration for H in HERE. In Fig. 6.9 the aspiration is visible for the synthetic version but not for the natural version. Furthermore, the peak of the fundamental frequency contour for stressed vowels occurs earlier in the natural speech than in the synthetic speech. This can be seen in Fig. 6.10 during the A of BOB.

Figure 6.9 Wideband spectrograms of synthetic and natural versions of "Larry and Bob are here".


7kHz
6kHz
5kHz
4kHz
3kHz
2kHz
1kHz
0kHz
Synthetic
Speech
LARRY
AND
BOB
ARE
HERE
VOICEPRINT LABORATORIES - P.O. BOX 218 - SOMERVILLE, NEW JERSEY

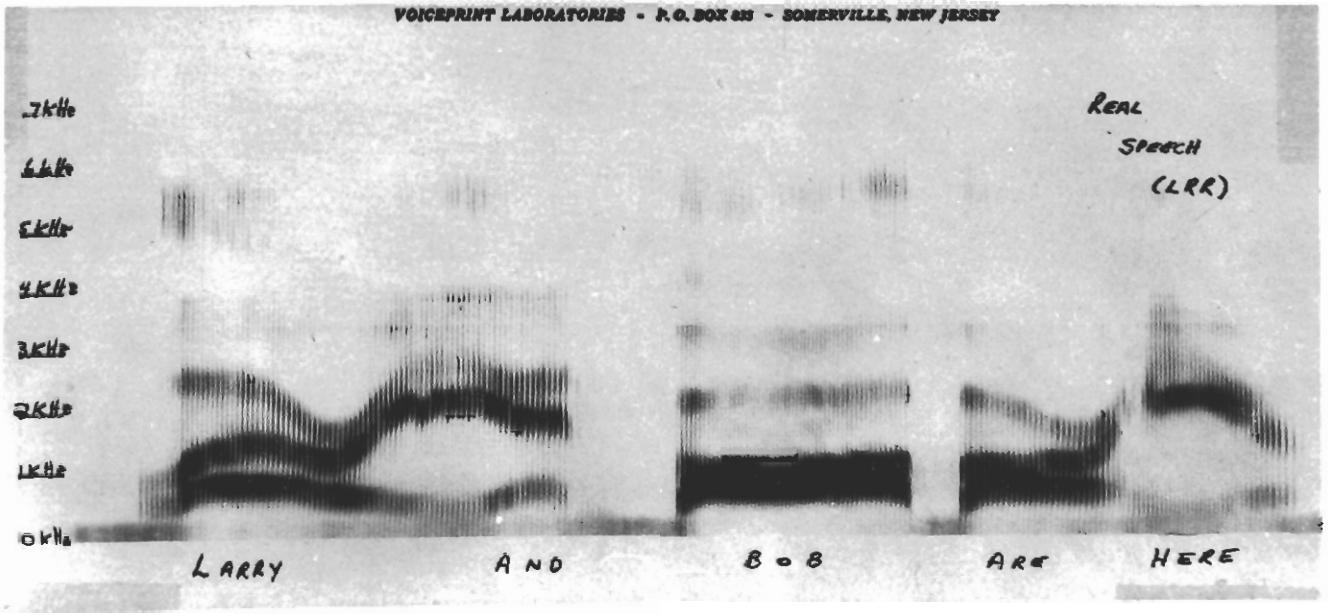
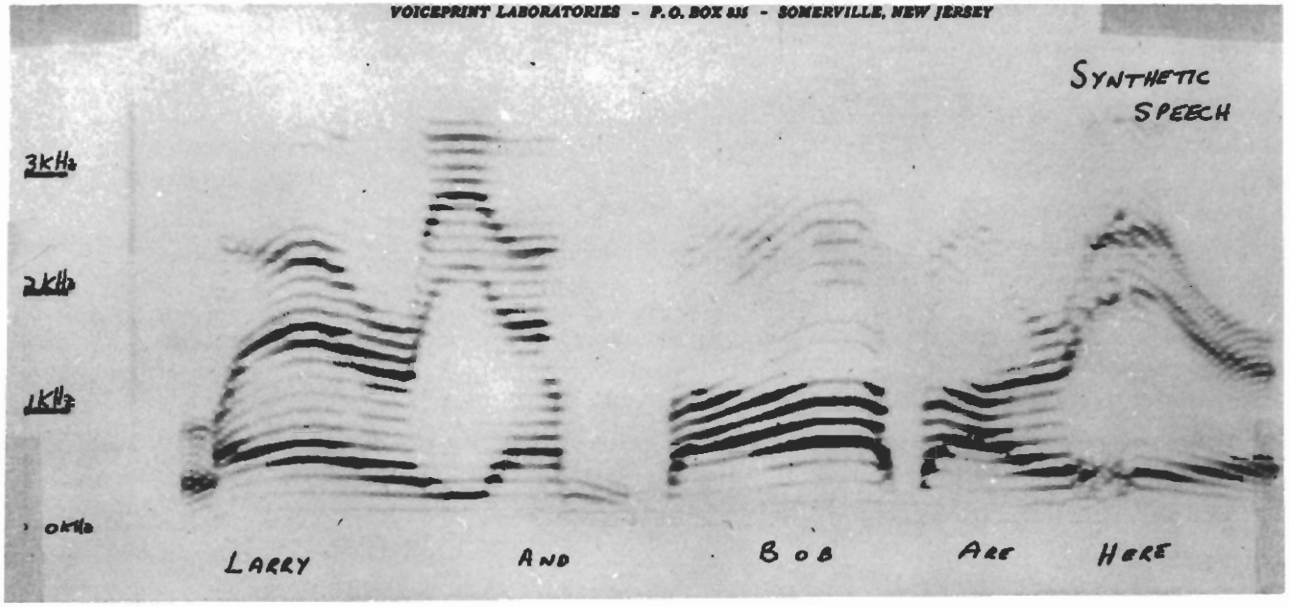
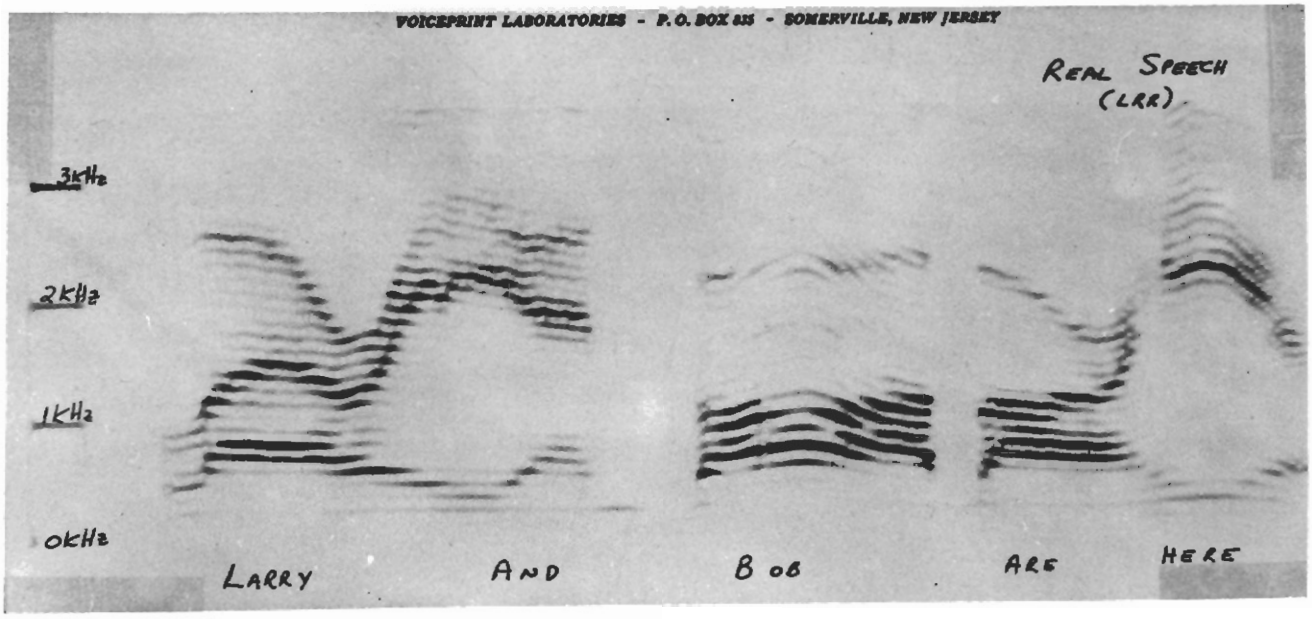

VOICEPRINT LABORATORIES - P.O. BOX 218 - SOMERVILLE, NEW JERSEY
7kHz
6kHz
5kHz
4kHz
3kHz
2kHz
1kHz
0kHz
REAL
SPEECH
(LRR)
LARRY
AND
BOB
ARE
HERE

Figure 6.10 Narrowband spectrograms of synthetic and natural versions of "Larry and Bob are here".


VOICEPRINT LABORATORIES - P. O. BOX 435 - SOMERVILLE, NEW JERSEY
SYNTHETIC SPEECH
3kHz
2kHz
1kHz
0kHz
LARRY
AND
BOB
ARE
HERE


VOICEPRINT LABORATORIES - P. O. BOX 435 - SOMERVILLE, NEW JERSEY
REAL SPEECH (LRR)
3kHz
2kHz
1kHz
0kHz
LARRY
AND
BOB
ARE
HERE

Figures 6.11 and 6.12 show wideband and narrowband spectrograms of synthetic and natural versions of the utterance "We sang all day". This utterance was unintelligible to our subjects. The responses to this utterance ranged from "We swam today" to "We _____ day". Only the first and last words were intelligible. An examination of Fig. 6.11 reveals why the middle portions of this utterance were unintelligible.

As seen in Fig. 6.11, the formant transitions from AE to NG in SANG are incorrect. In the synthetic speech formant two is not heading in the correct direction during this transition. In the synthetic version the presence of the nasal pole for NG (around 1 kHz) is manifested during the transition to NG and this is inappropriate. The spectrum of the synthetic speech during the NG does not match the spectrum of the natural speech.

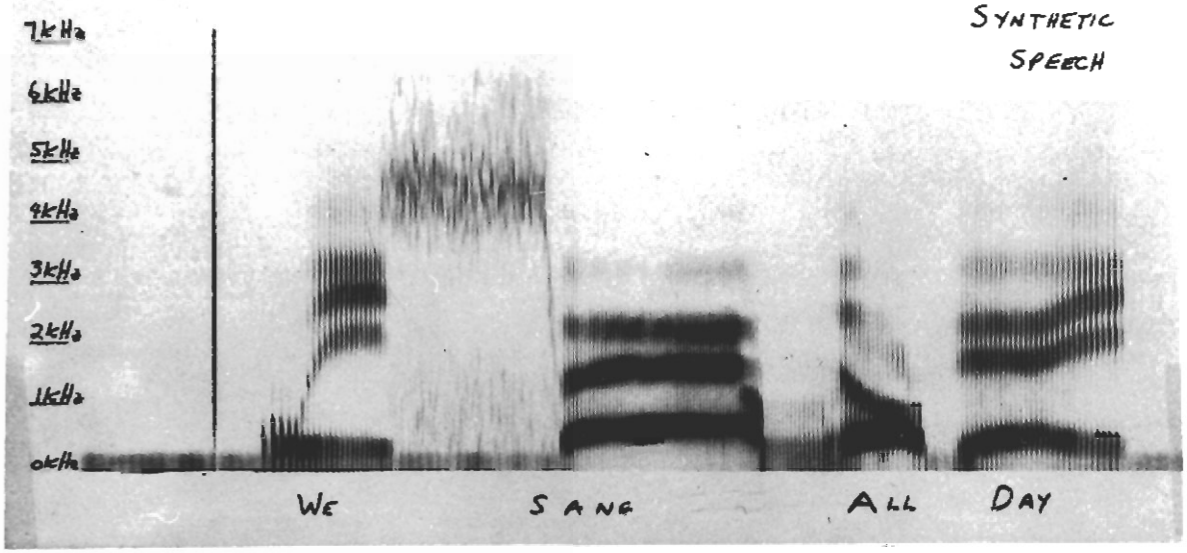
Another major error seen in Fig. 6.11 concerns the L in ALL. For the natural speech formant three is high during the L whereas it is quite low for the synthetic speech. Since the formant three target for L is 2575 Hz the reason for this discrepancy is unclear.

The fundamental frequency contours for the two versions of this utterance (Fig. 6.12) are similar in some ways. Both contours show peaks during the stressed vowel AE in SANG. Again the peak of the fundamental frequency contour for the synthetic speech occurs later in time than the peak for the natural speech. A rise in the FO contour during the voiceless consonant S of SANG is visible for both versions of this utterance.

The synthetic utterance "We sang all day" was unitelligible because two of the phonemes were poorly represented. In such a short utterance errors in phoneme characterization tend to make the utterance unintelligible and this was indeed the case here.

Figure 6.11 Wideband spectrograms of synthetic and natural versions of "We sang all day".

VOICEPRINT LABORATORIES - P.O. BOX 435 - SOMERVILLE, NEW JERSEY



VOICEPRINT LABORATORIES - P.O. BOX 435 - SOMERVILLE, NEW JERSEY

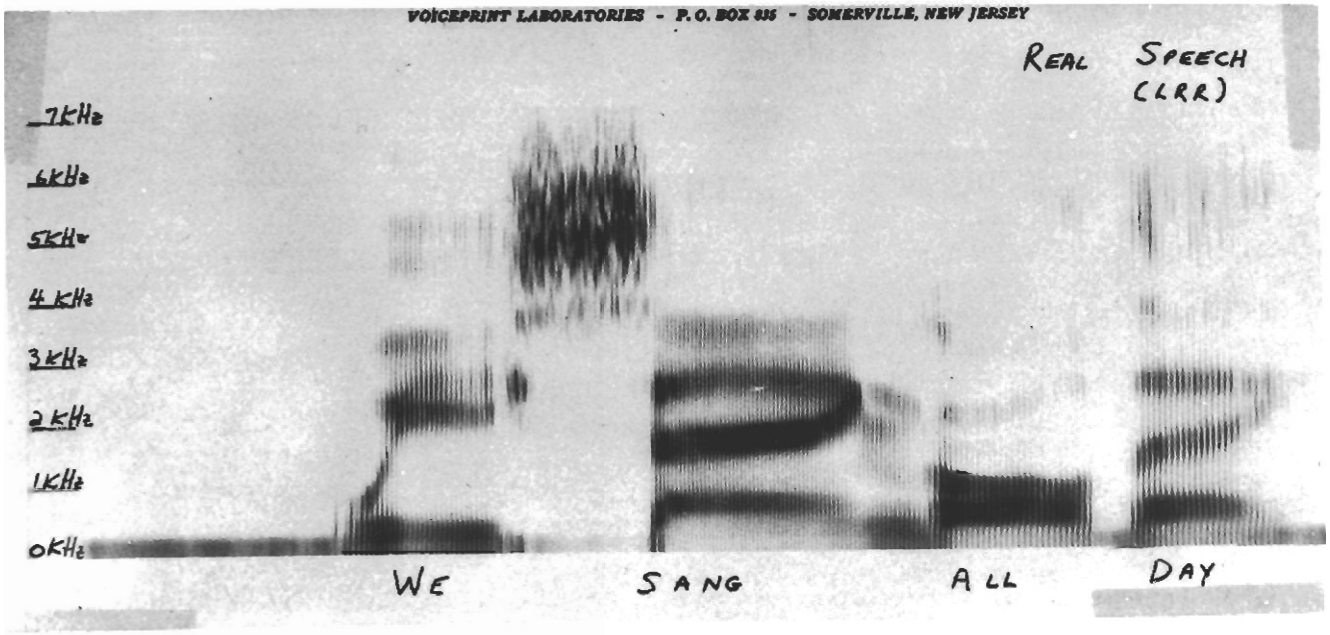
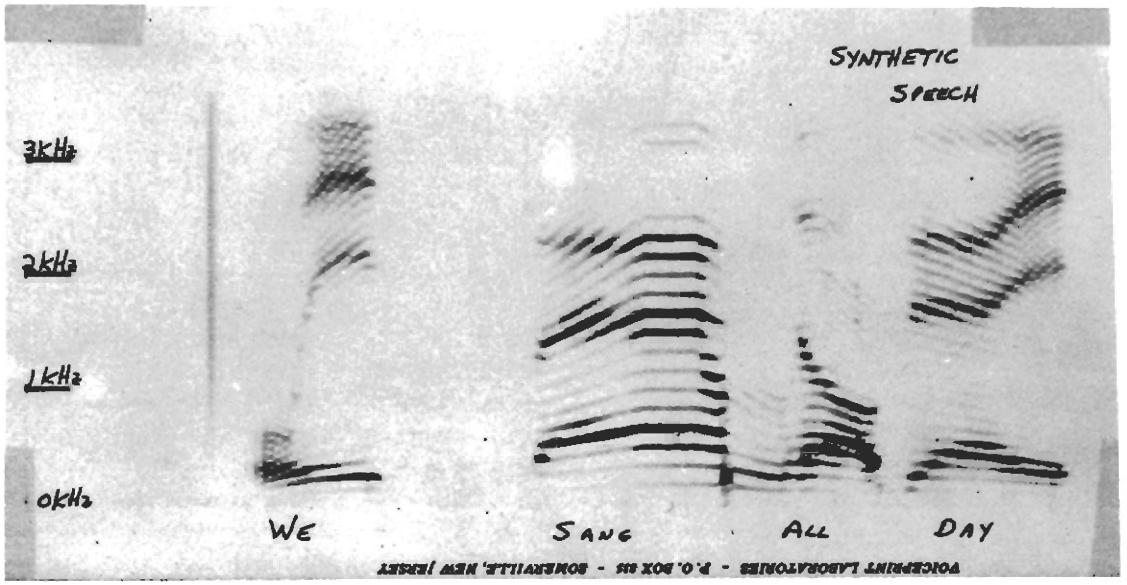
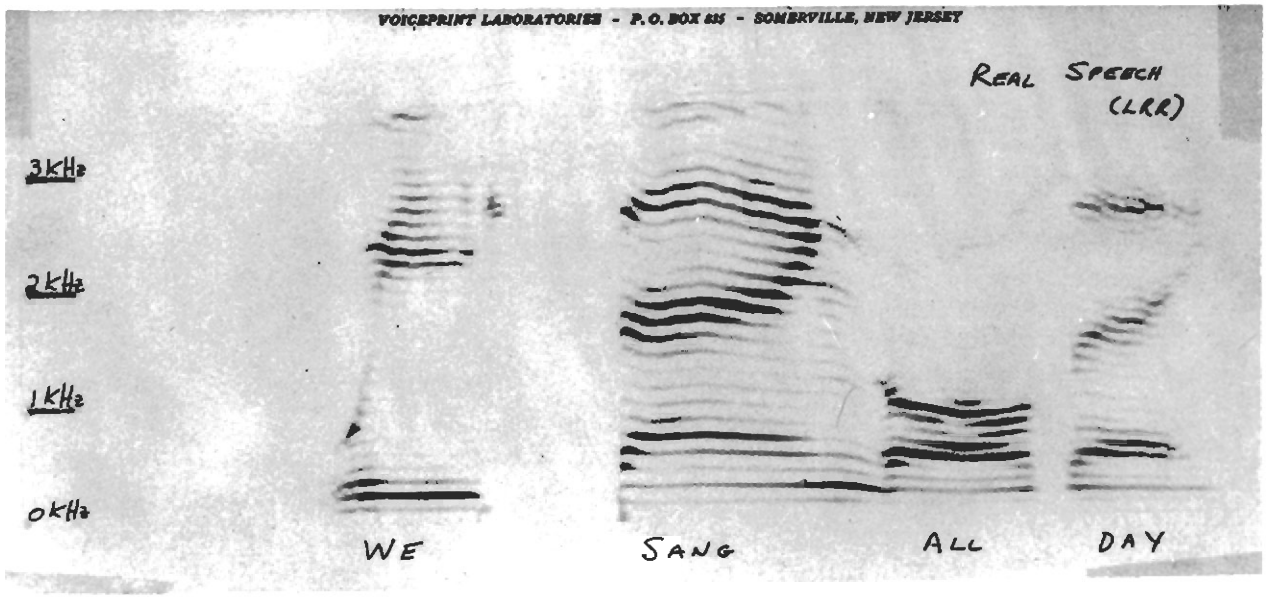


Figure 6.12 Narrowband spectrograms of synthetic and natural versions of "We sang all day".


SYNTHETIC
SPEECH
3kHz
2kHz
1kHz
0kHz
WE
SANG
ALL
DAY
VOICEPRINT LABORATORIES - P.O. BOX 435 - SOMERVILLE, NEW JERSEY


VOICEPRINT LABORATORIES - P.O. BOX 435 - SOMERVILLE, NEW JERSEY
REAL SPEECH
(LRR)
3kHz
2kHz
1kHz
0kHz
WE
SANG
ALL
DAY

CHAPTER SEVEN

DISCUSSION, PRACTICAL APPLICATIONS, AND CONCLUSIONS

7.1 Discussion of Results of Consonant Tests

In Chapter Six, we presented the results of our formal tests. For consonants in syllables the percent correct identification ranged in the seventies and low eighties. What do these scores mean, how do they compare with other results, and how can they be improved? Our goal in this section is to provide answers to these questions as best we can.

First of all, of what significance is a score of 80% correct intelligibility in our consonant evaluation tests? This score tells us that without the contextual information present in speech, consonants are properly identified four-fifths of the time. As we have seen, contextual information will raise these scores - often significantly. For our purposes, however, these scores served primarily as a basis of judging our rules for generating consonants. As such they have surely pointed out areas in which modifications were necessary - as for the nasals for example. They provided us with the quantitative basis necessary to test the effects of parameter changes.

How do our results compare with those obtained by other researchers? We have no direct basis of comparison as no other researcher working on synthesis by rule has run such tests. However there do exist guidelines from which we can judge our results. Researchers have run consonant perception tests using synthetic stimuli which were hand generated. Generally the scope of these tests was significantly smaller than ours - i.e., there were only 3 or 4 possible responses instead of

16 as in our test. Hence the probability of a correct guess was a significant factor in these tests. In these tests, generally, a single parameter was varied through a range of values to choose an optimum setting. The percentage of responses for a single consonant were plotted versus the parameter setting. The peak response percentages in such tests were 75-90 percent (Nakata, 1959, Heinz and Stevens, 1961 etc.). When the high probability of correct response for a guess is taken into consideration for these tests, we see that the peak percentages are comparable to those in our tests.

A further guideline for comparison is available from recent work of Strong (1967). Strong extracted the parameters for a parallel synthesizer from real speech by eye with the aid of a digital computer - which essentially provided a high degree of time and frequency resolution. He used as input a list of words beginning and ending in consonants. He synthesized these words, using this semi-automatic technique, and then asked listeners to identify the respective initial or final consonant. His list was not exhaustive in that not all consonants appeared in final and initial position. For initial consonants the overall correct percentage was 80% (ours was 73% with no fricative pooling), and for final consonants it was 86% (ours was 79% here). A major difference existed between the synthesizer Strong used (a parallel synthesizer) and the one we used. Thus it may be unfair to compare results but we are using the results only as a guideline. Strong used a totally different method to produce his parameters - he extracted them from real speech. This is significantly easier than generating them from phonetic symbols.

As to improving our results, we feel that this should certainly be possible. For the nasal consonants we will try some spectral matching techniques to determine optimal positions for the nasal pole and zero as well as optimal formant target positions. For G we will experiment with different bursts to see if a suitable one can be found. Finally, the target position of formant two for D and G preceding back vowels will be varied. Consonant intelligibility tests will be used to determine the effectiveness of the changes.

7.2 Discussion of Results of Sentence Tests

The tests with synthetic sentences have led to some definite results and observations. Among them are:

1. Simple sentences are highly intelligible using our scheme. There is no need for pauses, and errors in timing (vowel duration etc.) do not generally make the sentence unintelligible. Errors in phoneme characterization are of extreme importance.here.

2. Complex sentences are not quite as intelligible as simple sentences. One of the major difficulties here concerns timing as related to vowels. In our scheme stress on a vowel is signified in two ways. The duration is significantly lengthened and there is a wide peak in the fundamental frequency contour centered on the vowel. This combination has been found (by comparisons with stressed vowels of real speech) to be too much of an exaggeration of what happens in real speech. There the stressed vowels are characterized by either a peak in the fundamental frequency contour, or an increase in their duration, or, most generally, a combination of both these effects. However, when the combination of effects is used in real speech, the duration increase is smaller than when it is used alone, and similar results hold for the

peak in fundamental frequency contour. Our stressed vowels give the impression of always being unnaturally stressed - and usually unnaturally long. This often affects the intelligibility of the synthetic sentences. No simple solution to this problem has been found. It is a problem which deserves future consideration.

Furthermore, for the complex sentences, there is a strong need for rhythm to be introduced. The timing appears too uniform throughout the sentence. Modification of the fundamental frequency contour and/or the timing of events in producing the speech, based on the syntactic structure of the utterance, may perhaps provide the rhythm which is lacking in these sentences. Level variations throughout the utterance may also be of importance to the rhythm of a sentence. Until some understanding of the factors which determine rhythm is acquired, it will be difficult to provide the solution to our problem.

3. The fundamental frequency contour was generally felt to be a natural sounding contour which blended well with the speech. Questions were well reproduced using changes only in laryngeal tension to signify the question. The archetypal subglottal pressure contour provided reasonable sounding initial and terminal sections of our fundamental frequency contour. Stress was quite identifiable with our method. Subjects easily associated the peaks in our fundamental frequency contour with stressed vowels.

4. Every utterance was treated as a single breath-group. This was done due to lack of knowledge as to how to separate an utterance into more than one breath-group. However in many of our utterances, the need for two breath-groups was apparent. (For example the utterance

"A Bird In The Hand Is Worth Two In The Bush" is a natural candidate for two breath-groups when spoken with a pause after HAND. Rules will have to be devised to account for this situation.

5. Further acoustic correlates of word boundary are necessary for our scheme. Many of the errors in the sentences involved difficulty in finding a word boundary to separate words which were difficult to perceive. The words often tended to run onto each other causing whole clauses to be unintelligible. A prime target for this type of error was a clause having one word end in a consonant (say S) and having the next word begin in the same consonant. Listeners tended to place the consonant with the first word and tried to form a new second word without the consonant. One possible solution would be an intensity dip between the final and initial consonants. This intensity dip would be perceptually equivalent to a very short pause in the speech.

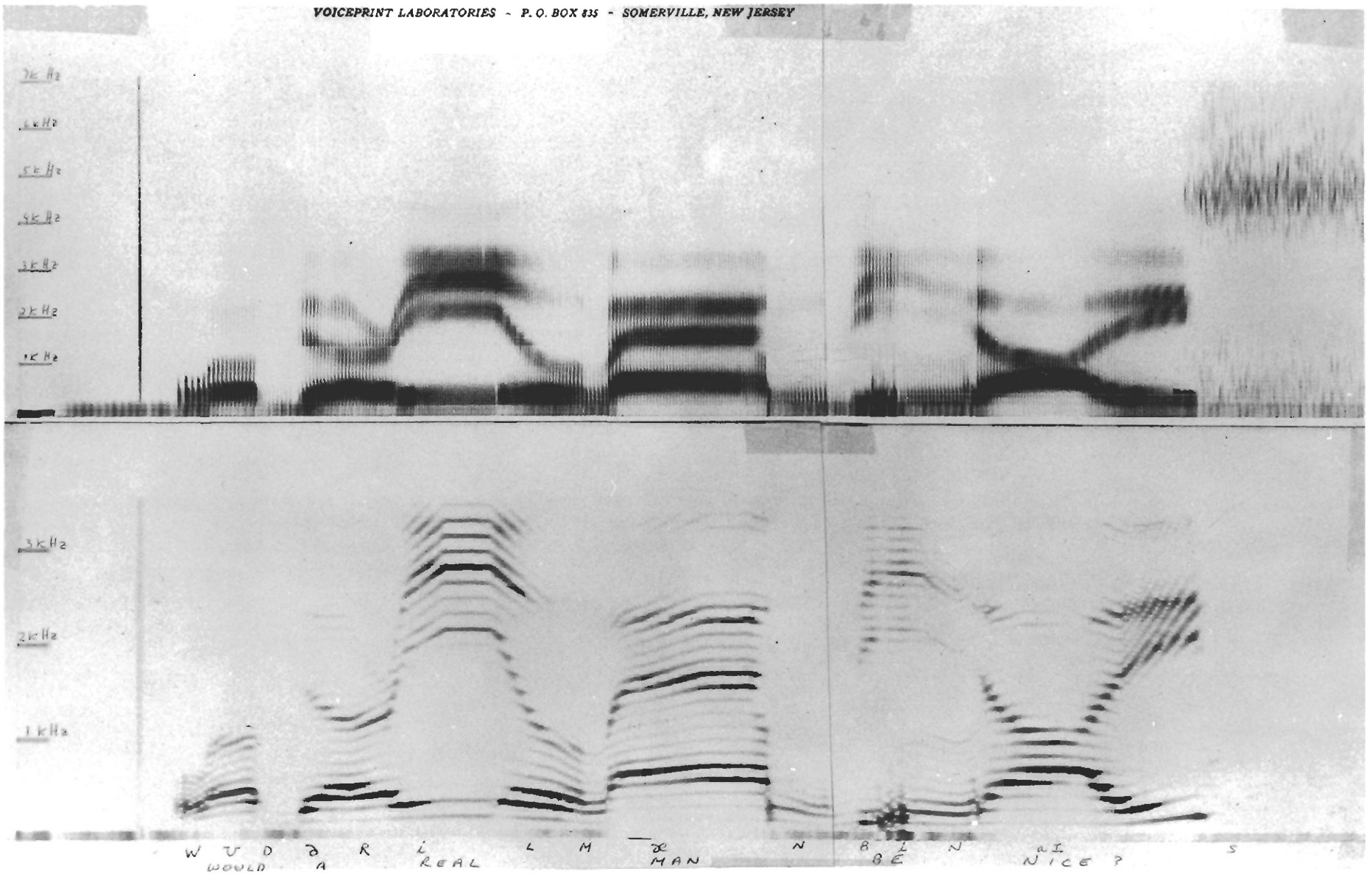
7.3 Further Examples of Synthesis

During the course of this thesis work a paragraph of connected speech, as well as our original group of 21 sentences, was generated for demonstration purposes. The paragraph was composed of nine very simple sentences written on the level of an elementary school student. The intelligibility of this paragraph was high, as judged by a group of five listeners. Both this paragraph and the group of 21 sentences are included on the demonstration tape which accompanies this thesis. A listing of the sentences and the paragraph is included in Appendix B.

An example of both wideband and narrowband spectrograms for a synthetic yes-no question is shown in Fig. 7.1. The utterance used was "Would a real man be nice?". The wideband spectrogram, in the upper half

Figure 7.1 Wideband and narrowband spectrograms of synthetic utterance "Would a real man be nice?".

VOICEPRINT LABORATORIES - P. O. BOX 835 - SOMERVILLE, NEW JERSEY



of Fig. 7.1, shows long steady state durations for the stressed vowels. These are the IY in REAL, the AE in MAN, and the A-IY in NICE. An examination of the narrowband spectrogram shows the peaks in the fundamental frequency contour during the stressed vowels, and the valleys of this contour during the voiced consonants. The fundamental frequency contour rises during the last 150 msec of voicing because this utterance is characterized by the laryngeal tension contour for the marked breath-group.

7.4 Practical Applications of Synthesis by Rule

One practical application of synthesis by rule is in a "reading machine" for the blind. A blind person would carry a box which contained an optical scanner, a transmitter and a receiver. The optical scanner, would scan the text to be read and send the information, suitably coded, to a main terminal nearby. The information would be fed into a computer which would determine the orthography of what was scanned. A program would then suitably convert the orthography into a set of phonemes, stress marks, and punctuation. The synthesis by rule program would then take control and produce the speech. This speech would then be sent back to the blind person and he would listen to it. There are many problems associated with this machine, and currently a good deal of work is being done trying to solve them. A solution to one of the problems, the conversion from orthography to phonemes, has been proposed and tested by Lee (1965). The method used by Lee is essentially a dictionary look-up procedure where phonetic transcriptions of a large number of root words are stored. Lee showed that with a reasonable size dictionary (on the order of 30,000 root words) and a set of combination and generation

laws, about 250,000 words of American English could be decomposed into phonemes.

A second application of this machine is a "talking machine" aid for someone who is unable to produce speech due to an accident, etc. He could key in the phonemic text on a small keyboard mounted on his body, and thus converse freely with another human. With an analog speech synthesizer, there would be no difficulties in converting from phonemes to control signals in real time.

Another application for synthesis by rule is in man-machine interaction. For on-line computation, having the computer "speak out" comments or answers, etc., would add an extra degree of freedom. The experimenter's eyes could be kept fixed on a display and need never turn to the console to check what is happening. The possibility of remote use of a computer by telephone is another potential application of a successful synthesis by rule scheme.

A final application is one related to telephony. One could send speech over channels of low information capacity since the rates associated with phoneme production are orders of magnitudes less than rates associated with the acoustic waveforms of speech.

7.5 Alternatives to Phoneme Synthesis

In this work we considered synthesis of speech from a string of phonemes. There are alternative methods of synthesizing speech.

One such method is speech synthesis from a library of stored sentences and words. The acoustic waveforms for a fixed list of several sentences are stored, either on magnetic tape, or in a computer memory. To synthesize speech, you abut the appropriate stored sentences in a serial fashion. This method would be appropriate for such places as the

weather bureau, or the stock exchange where the number of possible sentences needed is small. Thus, a few key sentences and a small list of additional words would be adequate for all situation. For instance, one key sentence for the weather bureau might be, "Good afternoon, this is the U. S. Weather Bureau forecast for Boston and vicinity. The present temperature is _____ degrees." For all situations, thirty additional words would be needed to describe the temperature, i.e., the word thirty-seven is the sum of the word thirty plus the word seven.

Problems associated with general use of such a scheme are clear. The amount of storage necessary to handle a large number of situations is enormous. There are also problems in abutting stored words and phrases. The intonation patterns must match at the boundaries, or there is an unnatural effect.

A somewhat different method of synthesizing speech is from a library of stored control signals for use with a terminal analog synthesizer. Speech would be synthesized by abutting the words with pauses between words. There are two major problems here. First, the amount of storage necessary to store the control signals for a large vocabulary, as would be needed to produce a large number of sentences, is enormous. Second, the speech is not connected and is therefore unpleasant to listen to for any length of time. (Try holding a conversation with someone, word by word.) If we try to combat the second problem by connecting the words, we have serious problems at the boundaries. We need rules to provide continuity of the control signals. The intonation pattern must be preserved across word boundaries. The problem with the necessary large storage still would not be solved.

7.6 Conclusions and Suggestions For Future Research

An acoustic domain approach to speech synthesis by rule has been formulated and programmed on a digital computer. Samples of speech have been generated using our scheme and their intelligibility has been measured. The problem of automatically generating fundamental frequency has also been worked on and a workable solution has been found.

The major successes and highlights of our scheme have been the following:

1. High consonant and vowel identifiability.
2. High intelligibility for simple sentences.
3. Reasonable to high intelligibility for complex sentences.
4. Faithful reproduction of fundamental frequency contours for both interrogative and declarative sentences.
5. A novel synthesizer design providing high quality voiced fricatives and provision for inclusion of a voicebar for voiced stops.
6. A novel method of handling formant transitions (our differential equation approach) and transition durations (our matrices of time constants).
7. Internal control of timing using system parameters with few external constraints.
8. A novel method of generating fundamental frequency data based on a physiological theory involving simulated laryngeal tension and subglottal pressure information.

As we terminate this thesis, the work is far from being completed. As there are many unsolved problems as well as improvements that can be made. These include:

1. A study of the diphthongs. Although treated as a two vowel sequence (with special constraints), the diphthongs turned out quite well. The diphthongs e^{I} and o^{U} still need further study to learn the effects of stress on these diphthongs. The diphthongs aI, I, and aU were well reproduced and need little further study. Using increased transition time to characterize a stressed diphthong turned out very successful. Formal tests should be run here to verify this point.

2. A study of the nasal consonants. The acoustic correlates of manner were poorly preserved by our scheme. Further study here will be necessary to isolate the factors which caused nasals to be heard as stops, fricatives, and W, L, or R.

3. A study of the affricates. We bypassed the problem of affricates as they occur very infrequently in speech. We treated them as two phoneme sequences. However there was little justification for our approach. An acoustic analysis of affricates is necessary before a good strategy for their synthesis can be devised.

4. A study of linguistic considerations in timing. A theory of vowel duration based on the semantic structure of the utterance is needed. A theory of timing as related to rhythm is necessary to discriminate between imperative, declarative, emphatic, and interrogative sentences. The same sentence can be said in more than one way and this should be accounted for acoustically.

This theory of timing would also have to insert pauses when necessary. Thus very long connected utterances could be synthesized using such a theory. This theory would have to determine when one breath-group ended and another began.

We feel we have made considerable progress in our acoustic domain approach. Just how far an acoustic domain approach can go is unclear. It is clear that situations exist when the vocal tract configuration is changing so rapidly that the concept of acoustic parameters is meaningless. (The vocal tract configuration is varying at a sufficiently high rate that representing it as a linear time invariant system is incorrect.) They cannot be measured due to the finite time it takes to make the measurement - during the measurement time, what you are trying to measure has changed significantly. Such situations exist primarily adjacent to periods of total closure such as the transitions of stops and nasals. During these periods our representation of the vocal tract is crude and we would expect problems. However solutions do exist - as evidenced by the high quality hand-synthesized speech attainable on a terminal analog synthesizer. It is this high quality that we are aiming to reach. We hope to continue research in this field until we reach our goals.

APPENDIX A

SYNTHESIZER PROGRAM

The BLØDI program for the synthesizer is included in the following pages. The following brief description of the boxes used should provide enough information for the average programmer to follow the program.

Each box in BLØDI has a name (the first set of columns), a function (the second set of columns), and a set of parameters, and outputs (the third set of columns). Many boxes have multiple inputs and hence the inputs are numbered. A computer value of 2^{18} is treated as a sample value of 1 in BLØDI - i.e., a parameter equal to 1/2 numerically is represented as 2^{17} in BLØDI.

In our synthesizer the following boxes are used:

1. SUB-(subtractor)-2 inputs, 1 output

   OUT=IN1-IN2

2. ADR-(adder)-2 inputs, 1 output

   OUT=IN1+IN2

3. MPR-(multiplier)-2 inputs, 1 output

   OUT=IN1 x IN2

4. AMP-(amplifier-1 input, 1 output, 2 parameters

   OUT=IN x 2^{-PAR2} x PAR1

   (In column 3, parameters are presented before outputs. Parameters defined by commas are zero, as are undefined parameters.)

5. BAT-(battery)-1 input, 1 output, 1 parameter

   OUT=IN+PAR

6. DEL-(delay)-1 input, 1 output, 1 parameter

   OUT=IN(delayed by PAR samples)

   (delay = 1 sample if no PAR is given.)

7. QNT-(quantizer)-1 input, 1 output, 2n parameters, where n=level of quantizer.

   PAR1=n, OUT=PAR2 for IN<PAR3

   OUT=PAR4 for IN<PAR5

   . .

   . .

   . .

   OUT=PAR(2n-2) for IN<PAR(2n-1)

   OUT=PAR2n for IN$\geq$PAR(2n-1)

8. DTS-(double throw switch)-3 inputs, 1 output, 1 parameter

   For IN1$\geq$PAR $\Rightarrow$ OUT-IN2

   IN1<PAR $\Rightarrow$ OUT=IN3

9. MAX-(maximum)-up to 4 inputs, 1 output

   OUT=MAX(IN1,IN2,IN3,IN4)

10. HLD-(hold)-2 inputs, 1 output, 1 parameter

    If IN2>PAR $\Rightarrow$ OUT=IN1

    IN2$\geq$PAR $\Rightarrow$ OUT=last output value

11. CNT-(counter)-1 input, 1 output, 5 parameters

$$OUT=\begin{cases} PAR3 & \text{every PAR1 times IN>PAR2} \\ PAR4 & \text{otherwise} \end{cases}$$

    Initially OUT=PAR3 when IN>PAR2 for only PAR1-PAR5 samples.

12. PLS-(pulser)-1 input, 1 output, 4 parameters

    If IN$\geq$PAR1 $\Rightarrow$ OUT=PAR3 for PAR2 samples

    If IN<PAR1 for PAR2 samples $\Rightarrow$ OUT=PAR4

13. WNG-(white noise generator)-1 output, 1 parameter

    OUT=sequence of uncorrelated samples from a Gaussian distribution with mean zero and standard deviation PAR, truncated at ±6 PAR.

14. INP-(input)-Information from outside enters program (from tape)

15. OUT-(output)-Information from program is put on tape

16. FMT-(formant filter)-3 inputs, 1 output

    OUT=Function of last three input samples

    IN2=Center frequency of complex pole

    IN3=Bandwidth of complex pole

17. ZER-(filter to produce a zero pair)-3 inputs, 1 output

    OUT=Function of last three input samples

    IN2=Center frequency of complex zero

    IN3=Bandwidth of complex zero.

The synthesizer program follows.

```
SUB22     SUB     Q22
Q22       QNT     2,,,524288,DTS22,DTS20/2,HF1/2,HF2/2,HF3/2
DTS22     DTS     2,B222
B222      BAT     1,DD22
DD22      DEL     DTS22/3,SUB22
DTS20     DTS     5,MPV/2
MPV       MPR     DEY
DEY       DEL     VSW,G32
VSW       MPR     F10
F10       FMT     F5
F5        FMT     F9
F9        FMT     F4
F4        FMT     F8
F8        FMT     F3
F3        FMT     F7
```

F7	FMT	F2
F2	FMT	F6
F6	FMT	F1
F1	FMT	GCR
GCR	AMP	1,-2,SHP
SHP	FMT	NPL
NPL	FMT	NZR
NZR	ZER	PLZ,ATØ
PLZ	DEL	ATØ/2
ATØ	SUB	ADR
G32	AMP	1,-2,SHP2
SHP2	FMT	LPF
		AM2
AM2	AMP	1,-4,M4
M4	MPR	AD3/2
AD3	ADR	ADR/2
LPF	FMT	BDN
BDN	BAT	-1024,MAX1
MAX1	MAX	DTS30/2
B10	BAT	1024,DTS30/3
DTS30	DTS	40,VFS/2
WNG	WNG	32768,VSN,DTS20/3
VSN	MPR	MPN/2
MPN	MPR	G2
G2	AMP	[redacted] 1,-10, UFS
VFS	MPR	FPL
FPL	FMT	FZR
FZR	ZER	G33
G33	AMP	1,-2,SHP1
SHP1	FMT	SP,DP
SP	DEL	DP/2
DP	SUB	AD3
ADR	ADR	MPØP/2
MPØP	MPR	AMP
AMP	AMP	1,-6,B27
B27	BAT	2048,ØUT

```
ØUT     ØUT     5,,480,,3
P1      PLS     499,40,0,1,MPV,MPN,MPØP
INP     INP     6,,480,B1,DO
B1      HLD     M4/2
DO      DEL     DØ1,HO
HO      HLD     P1
DØ1     DEL     BO
BO      BAT     -2048,D1,H1
D1      DEL     D2,H2
D2      DEL     D3,H3
D3      DEL     G4
G4      AMP     1,8,H4,D4
D4      DEL     H5,D5
D5      DEL     D6,H6
D6      DEL     D7,H7
D7      DEL     D8,H8
D8      DEL     D9,H9
D9      DEL     D10,H10
D10     DEL     G5
G5      AMP     1,-2,BC
BC      BAT     131072,H11,D11
D11     DEL     D12,H12
D12     DEL     D13,H13
D13     DEL     D14,H14
D14     DEL     D15,H15
D15     DEL     D16,H16
D16     DEL     D17,H17
D17     DEL     G6
G6      AMP     1,-6,H18
H18     HLD     SUB22/2
H17     HLD     F1/3,LPF/3
H16     HLD     F2/3
H15     HLD     F3/3
H14     HLD     NPL/3
H13     HLD     NZR/3
```

```
H12         HLD     FPL/3
H11         HLD     FZR/3
H10         HLD     1,HF1
H9          HLD     1,HF2
H8          HLD     1,HF3
H7          HLD     NPL/2
H6          HLD     NZR/2
H5          HLD     FPL/2
H4          HLD     FZR/2
H3          HLD     VSW/2,DTS30
H2          HLD     SN/2
H1          HLD     DTS20
F4B2        BAT     231362.,F4/2
F4B3        BAT     247988.,F4/3
F5B2        BAT     74837.,F5/2
F5B3        BAT     239862.,F5/3
F6B2        BAT     -76064.,F6/2
F6B3        BAT     226492.,F6/3
F7B2        BAT     -212557.,F7/2
F7B3        BAT     208667.,F7/3
F8B2        BAT     -304692.,F8/2
F8B3        BAT     177209.,F8/3
F9B2        BAT     -333070.,F9/2
F9B3        BAT     134480.,F9/3
F10B2       BAT     -246049.,F10/2
F10B3       BAT     58458.,F10/3
B2F0        BAT     503589.,SHP/2,SHP1/2,SHP2/2
B3F0        BAT     242483.,SHP/3,SHP1/3,SHP2/3
HF1         HLD     F1/2,LPF/2
HF2         HLD     F2/2
HF3         HLD     F3/2
C1          CNT     20,-1,2,,0,H1/2,H2/2,H3/2,H4/2,H5/2,H6/2,H7/2,
                    H8/2,H9/2,H10/2,H11/2,H12/2,H13/2,H14/2,H15/2,
                    H16/2,H17/2,H18/2,B1/2,H0/2
            END
NØ INPUT TØ BØX     C1
```

The zero (ZER) and formant (FMT) boxes are coded as follows:

```
ZER     LDQ*    6,4
        STQ     ZT
        LDQ*    7,4
        STQ*    6,4
        STØ*    7,4
        MPY*    1,4
        LLS     17
        RND
        STØ     ZT+1
        LDQ     ZT
        MPY*    2,4
        LLS     17
        RND
        SUB     ZT+1
        ADD*    7,4
        STØ     ZT+2
        CLS*    1,4
        ADD*    2,4
        ADD     =1.B17
        STØ     ZT+1
        CIA     ZT+2
        LRS     17
        DVP     ZT+1
        XCA
        TRA     10,4
ZT      BSS     3
FMT     STØ*    9,4
        SUB*    6,4
        XCA
        MPY*    2,4
        LLS     17
        RND
        ADD*    9,4
        STØ*    8,4
```

```
CLA*    7,4
STØ*    6,4
SUB*    9,4
XCA
MPY*    1,4
LLS     17
RND
ADD*    8,4
STØ*    7,4
TRA     10,4
```

APPENDIX B

INPUT SENTENCES

Paragraph in demonstration tape.

This is a story about a man. The man works in a factory. What is his name? The man's name is Bob. What does Bob do? Bob helps make steel on a big machine. Bob works from nine to five. Then he goes home to his wife and family. Bob enjoys his life.

Twenty-one standard sentences in demonstration tape.

1. We were away a year ago.
2. I need a transistor radio.
3. I love you.
4. They drank a coke with rum therein.
5. I am speaking dumb and vain words.
6. Sue the bank under a false name.
7. The climb was warm and done without water.
8. She lay prone and hardly moved a limb.
9. Would a real man be nice?
10. Are you a red or yellow cat?
11. There are more than two factors here.
12. Noon is the sleepy time of day.
13. Welcome to this meeting.
14. Larry Rabiner is here.
15. My aptitude is floating.
16. This is an example of speech.
17. My name is Larry.
18. A bird in the hand is worth two in the bush.

19. You are a real liar.

20. Men strive but seldom get rich.

21. Are you a good boy or a bad boy?

REFERENCES

Beranek, L. (1949) Acoustic Measurements, John Wiley and Sons Inc., New York, 773-777.

Bruce, R. (1966) An investigation of American English stop consonants, M.S. Thesis, Mass. Inst. of Tech.

Cooper, F., Liberman, A., Lisker, L. and Gaitenby, J. (1962) Speech Synthesis by Rules. Speech Communication Seminar, Stockholm, August 29 to September 1, 1962. Paper F2.

Delattre, P. (1958) Acoustic Cues in Speech. Phonetica 2, 108-118, 226-251.

Delattre, P., Liberman, A., and Cooper, F. (1955) Acoustic loci and transitional cues for consonants. J. Acoust. Soc. Am. 27, 769-773.

Denes, P. (1963) On the statistics of Spoken English. J. Acoust. Soc. Am. 35, 892-904.

Dunn, H. (1961) Methods of Measuring Vowel Formant Bandwidths. J. Acoust. Soc. Am. 33, 1737-1746.

Flanagan, J. L. (1965) Speech Analysis Synthesis and Perception. New York: Academic Press, Inc.

Flanagan, J. L., Coker, C. and Bird, C. (1963) Digital Computer Simulation of a Formant-Vocoder Speech Synthesizer. Paper presented at the 15th Meeting of the Audio Engineering Society.

Fujimura, O. (1961) Some Synthesis Experiments on Stop Consonants in the Initial Position. Quarterly Progress Report No. 61, Research Lab. of Electronics, M.I.T., April 1961, 153-162.

Fujimura, O. (1962) Analysis of Nasal Consonants. J. Acoust. Soc. Am. 34, 1865-1875.

Guelke, R. and Smith, E. (1963) Distribution of Information in Stop Consonants. Proc. I. E. E. 110, 680-689.

Halle, M. (1964) On the Bases of Phonology. In J. A. Fodor and J. J. Katz (eds) The Structure of Language, Prentice-Hall: Englewood Cliffs, New Jersey, pp. 324-353.

Heinz, J. and Stevens, K. (1961) On the Properties of Voiceless Fricative Consonants. J. Acoust. Soc. Am. 33, 589-596.

Henke, W. (1966) Dynamic Articulatory Model of Speech Production Using Computer Simulation. Ph.D. Thesis, Mass. Inst. of Tech.

Holbrook, A. and Fairbanks, G. (1962) Diphthong Formants and Their Movements. J. Speech and Hearing Research, 5, 38-58.

Holmes, J., Mattingly, I. and Shearme, J. (1964) Speech Synthesis by Rule. Language and Speech, 7, 127-143.

House, A. (1961) On Vowel Duration in English. J. Acoust. Soc. Am. 33, 1174-1178.

Hughes, G. and Halle, M. (1956) Spectral Properties of Fricative Consonants. J. Acoust. Soc. Am. 28, 303-310.

Kelly, J. and Vyssotsky, V. (1961) A Block Diagram Compiler. Bell System Tech. J. 40, 669-676.

Kelly, J. and Lochbaum, C. (1962) Speech Synthesis. Speech Communication Seminar Stockholm, August 29, to September 1, 1962. Paper F7.

Kelly, J. and Gerstman, L. (1961) An Artificial Talker Driven From a Phonetic Input. (Abstract) J. Acoust. Soc. Am. 33, 835.

Kelly, J. and Gerstman, L. (1964) Synthesis of Speech from Dode Signals. U.S. Patent #3,158,685.

Kenyon, J. and Knott, T. (1953) A Pronouncing Dictionary of American English. G. and C. Merriam Co., Springfield, Mass.

Kim, C. W. (1966) The Linguistic Specification of Speech. Ph.D. Thesis, Univ. of the City of Los Angeles.

Klatt, D. (1967) Articulatory Activity and Air Flow During the Production of Fricative Consonants. Quarterly Progress Report No. 84, Research Lab. of Electronics, M.I.T., Jan. 1967, 257-260.

Lee, F. (1965) A Study of Grapheme to Phoneme Translation of English. Ph.D. Thesis, Mass. Inst. of Technology.

Lehiste, I. (1959) An Acoustic-Phonetic Study of Internal Open Juncture. Univ. Of Mich. Speech Research Lab., Report #2.

Lehiste, I. (1962) Acoustical Characteristics of Selected English Consonants. Univ. of Mich. Speech Research Lab., Report #9.

Lehiste, I. and Peterson, G. (1960) Studies of Syllable Nuclei 2. Univ. of Mich. Speech Research Lab., Report #4.

Lehiste, I. and Peterson, G. (1961) Transitions, Glides, and Diphthongs. J. Acoust. Soc. Am. 33, 268-277.

Liberman, A., Delattre, P. and Cooper, F. (1952) The Role of Selected Stimulus Variables in the Perception of the Unvoiced Stop Consonants. Amer. J. of Psych. 65, 497-516.

Liberman, A., Delattre, P., Cooper, F. and Gerstman, L. (1954) The Role of Consonant-Vowel Transitions in the Perception of the Stop and Nasal Consonants. Psychol. Monogr. 68, No. 8 (Whole No. 379).

Liberman, A., Delattre, P., Gerstman, L. and Cooper, F. (1956) Tempo of Frequency Change as a Cue for Distinguishing Classes of Speech Sounds. J. of Experimental Psych. 52, 127-137.

Liberman, A., Ingemann, F., Lisker, L., Delattre, P. and Cooper, F. (1959) Minimal Rules for Synthesizing Speech. J. Acoust. Soc. Am. 31, 1490-1499.

Lieberman, P. (1966) Intonation Perception and Language. Ph.D. Thesis, Mass. Inst. of Tech.

Mattingly, I. (1966) Synthesis by Rule of Prosodic Features. Lang. and Speech 9, 1-13.

Maxey, H. (1963) Terminal-Analog Synthesis of Voiced Fricatives. (Abstract) J. Acoust. Soc. Am. 35, 1890.

Nakata, K. (1959) Synthesis of Nasal Consonants by a Terminal Analog Synthesizer. Jour. of the Radio Research Laboratories 6, 243-254.

Nakata, K. (1959) Synthesis and Perception of Nasal Consonants. J. Acoust. Soc. Am. 31, 661-666.

Nakata, K. (1960) Synthesis and Perception of Japanese Fricative Sounds. Jour. of the Radio Research Laboratories 7, 319-333.

Ohman, S. (1966) Coarticulation in VCV Utterances: Spectrographic Measurements. J. Acoust. Soc. Am. 39, 151-168.

Peterson, G. and Barney, H. (1952) Control Methods Used in a Study of the Vowels. J. Acoust. Soc. Am. 24, 175-184.

Peterson, G. and Lehiste, I. (1960) Duration of Syllabic Nuclei in English. J. Acoust. Soc. Am. 32, 693-703.

Peterson, G., Wang, W., and Sivertsen, E. (1958) Segmentation Techniques in Speech Synthesis. J. Acoust. Soc. Am. 30, 739-742.

Rabiner, L. R. (1966) Speech Synthesis by State Simulation. (Abstract) J. Acoust. Soc. Am. 40, 1272.

Rader, C. and Gold, B. (1965) Digital Filter Design Techniques. Unpublished Report of Lincoln Laboratory.

Rosen, G. (1958) A Dynamic Analog Speech Synthesizer. J. Acoust. Soc. Am. 34, 201-209.

Stevens, K., House, A. and Paul, A. (1966) Acoustical Description of Syllabic Nuclei: An Interpretation in Terms of a Dynamic Model of Articulation. J. Acoust. Soc. Am. 40, 123-132

Stevens, K. (1966) On the Relations Between Speech Movements and Speech Perception. Paper presented at the XVIIIth International Congress of Psychology, Moscow. August 1-7, 1966.

Strong, W. (1967) Machine-Aided Formant Determination for Speech Synthesis. Unpublished Report, Air Force Cambridge Research Laboratories.

Wang, W. and Peterson,G. (1958) Segment Inventory for Speech Synthesis. J. Acoust. Soc. Am. 30, 743-746.

BIOGRAPHY

Lawrence Richard Rabiner was born on September 28, 1943 in Brooklyn, New York. He attended G. W. Wingate H. S. and graduated in June, 1960.

As an undergraduate at the Massachusetts Institute of Technology, he participated in the Cooperative Course in Electrical Engineering in conjunction with Bell Telephone Laboratories. He received simultaneous S.B. and S.M. degrees in June, 1964, after submission of a thesis entitled: Binaural Masking: The Effects of Interaural Delay of the Noise on the Detection of Tones.

He continued his graduate studies at M.I.T. During the period September, 1964, to June, 1967, he held a National Science Foundation Fellowship.

During the summers of 1965 and 1966 he was hired by the Bell Telephone Laboratories to do research on speech synthesis and speech intelligibility in noise.

He taught courses in introductory electrical engineering, and signal and system theory during the academic year beginning September, 1966.

He has been elected to membership in Eta Kappa Nu, Tau Beta Pi, and Sigma Xi. He is also a member of the Acoustical Society of America.

He is the coauthor, with N. I. Durlach and C. L. Laurence, of the paper: Further results on binaural unmasking and the EC model. This paper appeared in the Journal of the Acoustical Society of America, Volume 40, July, 1966, pp. 62-71.